Investigating algorithmic bias mitigation in the public sector

Hadley Beresford

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Sheffield Methods Institute

October 2023

Abstract

This thesis makes an original contribution to critical algorithm studies by addressing the gap in the literature regarding the experiences and perceptions of data practitioners utilising algorithmic bias mitigation methods. This is important because while algorithmic bias mitigation methods have been proposed, little is known about how data practitioners engage with them, nor how practitioners' perceptions regarding these methods may impact their effectiveness. Understanding such things is crucial, as how data practitioners engage with these methods may have implications for the effectiveness of algorithmic bias mitigation efforts within an organisational context.

The thesis makes its contribution through three empirical qualitative papers, which together aim to investigate how practitioners in a government department might work to mitigate the impact of algorithmic bias. The research was carried out in partnership with the Department of Work and Pensions (DWP), the UK's ministerial department responsible for implementing work and welfare services and policy.

The first paper reported on research that used semi-structured interviews to investigate how data practitioners at DWP are engaging with algorithmic bias mitigation methods. The second paper investigated how practitioners on the Aurora AI project, a Finnish AI recommender project run by the Finnish Ministry of Finance, were working towards 'good practice' in algorithmic bias mitigation. This research also used semi-structured interview methods, interviewing Aurora AI team members and AI Ethics Experts. The third paper reports on research that used workshop methods to investigate how DWP organisational culture might influence the adoption of mitigation approaches.

Through analysis of the findings of these empirical chapters, this thesis makes three overarching contributions to knowledge. The first is that the fast-paced working practices that characterise the development of algorithmic technologies is not conducive to the slower-paced thinking needed to consider algorithmic bias using a socio-technical lens. Often, practitioners are under pressure to produce results quickly, and this may lead to the prioritisation of immediately tangible results such as a project's technical deliverables. The second contribution is to highlight the importance of context and, in my case, the significance of the UK civil service context and the unique challenges which exist therein. Specifically, algorithmic technologies deployed within a civil service context are strongly influenced by political processes and build on policy decisions already put in place by government officials. Finally, due to these practitioners' position as civil servants, they may be required to consider the diverse and conflicting views in found in the public in a way private organisations do not. However, the views of the public are currently missing from discussions of how the public sector should engage with algorithmic technologies, leaving practitioners to imagine what the publics views might be.

In addition to contributions to the emerging fields of critical algorithm and data studies, this thesis contributes towards a range of disciplines interested in the role of algorithmic technologies in society, including established fields such as information studies, sociology, communication studies and organisation studies.

Table of Contents

1.1. The rise of algorithmic technologies

While the public imagination had been caught by AI in the early 1980s, with films such as Blade Runner, Terminator, and Tron, hopes that anything practically useful would come out of AI had dwindled by the late 1980s (Bostrom, 2014). From the late 1980s through to the early 1990s, interest in Artificial Intelligence (AI) was fairly minimal (Bostrom, 2014; Newquist, 2020). By this time, the commercial success of 'expert systems', an early AI system designed to replicate human experts' decision making, had failed to materialise – along with the technology's promised benefits to efficiency and productivity (Bostrom, 2014). AI programmes were being defunded, and academics started avoiding the term AI to describe their work in funding applications, so as not to be associated with the then maligned technology (Newquist, 2020).

Much has changed in recent years with the rise of 'algorithmic technologies', a type of AI suited to performing specific tasks such as image recognition or matching customers to adverts (also known as narrow AI). Some have stated these technologies are economically important, with PwC estimating AI could contribute $15.7 trillion to the world's economy by 2030 (PricewaterhouseCoopers, 2017). The UK government has invested £2.3 billion into AI since 2014, across a range of different initiatives including the NHS and postgraduate education (Office for Artificial Intelligence, 2021). Public sector departments have increasingly been making use of algorithmic technologies, and according to an investigation by *The Guardian*, 140 out of 408 councils have invested in algorithmic software packages (Big Brother Watch, 2021; Marsh and McIntyre, 2020). Additionally, The UK's National AI Strategy and the UK's Innovation Strategy both position AI and algorithmic technologies as being central to the UK's economic development plans, due to their potential for efficiency gains and breakthrough discoveries (Office for Artificial Intelligence, 2021)

In the context of this increasing use of algorithmic technologies, there is concern that these technologies discriminate against marginalised groups; a phenomenon often referred to as 'algorithmic bias'. Concerns about algorithmic bias have been particularly prominent within the public sector, because of the harmful repercussions which have arisen as a result of the deployment of these algorithms therein. This differs from their use in a private sector context, because the public sector often provides essential services to people, such as healthcare or welfare. The perils of algorithmic bias have therefore been highlighted in several public sector service contexts. For example, algorithms used in the US criminal justice system have been found to give black defendants higher reoffending risk ratings than white defendants with a similar criminal history (Angwin *et al*., 2016; Kirkpatrick, 2016). Algorithms used in child protection services in the US, which aim to detect the likelihood of children being abused or neglected, have been found to discriminate against families with more extensive experiences of poverty than those without (Eubanks, 2018). These harms have been recognized within global and regional policy, such as in the OCED's AI Policy Principles (Yeung, 2020), as well as the AI strategies of the EU and UK government, which state that it is important that these

new technologies do not embed old biases and are not discriminatory (Balayn and Gürses, 2021; Office for Artificial Intelligence, 2021).

Despite these concerns, algorithmic technologies are being adopted in UK public sector practice. In 2017, Durham police started using an algorithmic technology to assist with police work called HART (Harm Assessment Risk Tool). This tool was designed to classify offenders depending on their perceived likelihood of reoffending. These classifications were then used to aid in decisions about whether offenders would be offered access to a rehabilitation programme for offenders at low-risk of reoffending. However, there were concerns the HART system would discriminate against marginalised communities (Oswald *et al.*, 2018). The tool was specifically designed to limit the number of false negatives produced, to ensure those who committed serious crimes were not given a chance to reoffend (Big Brother Watch Team, 2018). However, due to the decision to limit false negatives, the algorithm would over-estimate likelihood of reoffending, and in doing so was more likely to discriminate against marginalised groups. In another well-known example, due to social distancing brought about by the Covid-19 pandemic, UK A-Level students were unable to sit their final exams. In place of exams, teachers' estimated grades for each individual student were used, in conjunction with the school's performance in previous years, to produce the students' grade. Due to the use of school's past performance data, students who attended state schools were more likely to have their teacher-estimated grade downgraded, whereas students who attended selective private schools were more likely to receive their estimated grade without downgrading (Clement-Jones, 2021).

However, algorithmic technologies are still increasingly being relied on to drive efficiency gains across the public sector and are now used to target police efforts, provide clinical insights to the NHS, and answer customer queries within government services (Dencik *et al.*, 2018; Oswald *et al.*, 2018; Hughes, 2019). In 2019, it was announced by the Digital Chief of the DWP (Department of Work and Pensions) Simon McKennin, that data and artificial intelligence would be top priorities for the department in the year ahead (Trendall, 2019). In 2021, DWP started trialling an algorithm which detects fraud in Universal Credit claims, with plans to make the algorithm prevent payment of fraudulent claims before they are paid out (Public Law Project, 2022). Recently, DWP committed £70m worth of investments towards their digital transformation fund to expand their use of algorithmic technologies, some of which is expected to go towards fraud prevention (Waterfield, 2023; DWP, 2023). In a related move, DWP have committed to generating £1.3 billion worth of savings through their counter-fraud activity in 2023-2024 (DWP, 2023).

It has been argued that economic factors play a considerable role in the motivation to adopt algorithmic technologies. The Data Justice Lab's report *Investigating uses of citizen scoring in public services* suggests one of the key drivers of algorithmic technologies in the UK has been the impact of austerity – councils are having to do a lot more, with a lot less (Dencik *et al.*, 2018). Indeed, Eubanks (2018) notes that the increase of algorithmic technologies in public services has occurred alongside rapidly rising economic insecurity in the last decade. Additionally, she posits that algorithmic technologies are shaped by the fear of economic insecurity and in turn shape the politics and experiences of those in poverty (Eubanks, 2018).

Within the UK context, these concerns come at a time where the UN Rapporteur has said "[t]he British welfare state is gradually disappearing behind a webpage and an algorithm, with significant implications for those living in poverty." (Alston, 2019, p13). These insights, along with the previously discussed developments seen at organisations such as the DWP, highlight a tendency for algorithmic technologies to be utilised where the target population may already be quite vulnerable, such as those requiring council or government assistance.

Academics have argued that the very nature of public sector work – being political and requiring keen judgement to assess and evaluate information which cannot be quantified in a straightforward manner – makes automated approaches ill-suited to the types of tasks in the public sector (Veale and Brass, 2019; Northrop *et al*. 1990). Furthermore, the population subject to the decisions of these algorithms are particularly vulnerable. If they are discriminated against, they have less recourse to act against it. Because of these concerns, it is important that if these algorithmic technologies are to be developed, then some efforts are put in place to mitigate against the risks posed by algorithmic bias.

The context surrounding the adoption of algorithmic technologies, and algorithmic technologies' propensity towards producing biased outputs, underscores the need for research on algorithmic bias mitigation approaches in the public sector. The current environment in which these technologies are adopted means the most vulnerable are often those who are subject to the decisions made by algorithmic technologies, presenting risk to those already marginalised. In response, there is growing interest across the UK to assess the potential risks of algorithmic bias in addition to identifying algorithmic bias mitigation practices. Initiatives include the investigation by the UK government's Centre for Data Ethics and Innovation (CDEI) and the Cabinet Office's Race Disparity Unit into algorithmic bias (CDEI, 2020). Additionally, the CDEI have released their Data Ethics Framework, which contains non-legal guidance about how data should be used in the public sector (UK Government, 2018; Veale and Brass, 2020). The Ada Lovelace Institute have produced a report detailing potential approaches for mitigating algorithmic harm in an NHS context (Ada Lovelace Institute, 2022). Alongside these national level responses, international academic communities, such as members of the FATE (Fairness, Accountability, Transparency, and Ethics) community, have discussed and interrogated frameworks for understanding and mitigating algorithmic bias (ACM FAT*, 2019).

To mitigate the risks of algorithmic bias, data scientists have developed algorithmic de-biasing methods, using quantitative techniques that rely on data scientists producing and assessing output metrics such as error rates and other comparative statistics (Balayn and Gürses, 2021). However, it has been noted by prominent scholars that further research in this area needs to include qualitative research that can capture "the messy reality of many contemporary on-the-ground situations" (Veale and Binns, 2017, p12).

To date, there has been little qualitative research regarding data science practitioners' perceptions and experiences of algorithmic bias and algorithmic bias mitigation methods. Orr and Davies (2020), Veale *et al*. (2018), and Holstein (2017) have interviewed practitioners to understand how they are situated within the development of algorithmic bias, and their

responsibilities and engagements within their working context. However, their analysis is primarily focused on how individual actors are constrained within a collective system (Holstein, McLaren and Aleven, 2017; Veale, Van Kleek and Binns, 2018; Orr and Davis, 2020). Conversely, the way in which algorithmic bias mitigation is approached from an organizational or project perspective has received less attention. Additionally, while Veale *et al.*'s (2018) study investigates the challenges faced by public sector practitioners in mitigating algorithmic bias, although the findings of this are limited to a US context.

Moreover, despite the growing activity seen in the public, private, and academic sectors around algorithmic bias, there has been little research on how civil servants in UK organisations make sense of these algorithmic bias mitigation efforts and how organisational processes might mediate these efforts. Without knowledge of how algorithmic bias mitigation efforts are understood, engaged with, and how they inform working practices, it is not possible to know how effective these mitigation methods might be in practice. In addition, little is known about how civil servants experience attempts to implement algorithmic bias mitigation methods and what challenges they might face.

This thesis makes an original contribution to critical algorithm studies by addressing the gap in the literature regarding the experiences and perceptions of data practitioners utilising algorithmic bias mitigation methods. This is important because while algorithmic bias mitigation methods have been proposed, little is known about how data practitioners engage with them, nor how practitioners' perceptions of these methods may impact their effectiveness. As discussed earlier, the UK civil service department the DWP is looking to adopt more algorithmic technologies. Thus, it is a pressing matter that the issue of algorithmic bias within the DWP is investigated and that algorithmic bias mitigation practices are established to mitigate the risks of further marginalising marginalised groups. Furthermore, there is currently an absence of empirical investigation of efforts to implement bias mitigation in a UK civil service context, and the organisational factors that help or hinder these bias mitigation efforts. This leads to the aims and objectives of the research papers which constitute my thesis, which focus on algorithmic bias mitigation practices in a public sector context. The following section describes the aims and objectives of my thesis papers. In view of the issues discussed in this section, it is a matter of some urgency that the issue of algorithmic bias within the DWP is investigated. In the following section, I detail my approach to research design in the DWP context.

1.2. Aims, objectives, and research questions overview

The DWP is responsible for welfare and social security payments within the UK and is the largest UK civil service department by expenditure. In response to its own increasing concern of the risks of algorithmic bias, the Department of Work and Pensions (DWP) has sought to develop working practices to mitigate against the risks of algorithmic bias, including supporting my PhD project exploring algorithmic bias in the DWP and ways of mitigating it.

As part of this collaboration, DWP provided a short brief for the PhD project. The brief for the PhD project was written in collaboration between contacts at DWP and my supervisors Professor Helen Kennedy and Professor Jo Bates. At the time the brief was written, my supervisors were already collaborating with the DWP on another research project, and identified algorithmic bias as an upcoming concern for the department which warranted further research. Funding for the PhD project came from the Centre for Doctoral Training (CDT) for Data Analytics and Society, an Economic and Social Research Council (ESRC) funded CDT focused on bringing together the social sciences and advanced quantitative methods, and evaluating the new roles played by data in society. Furthermore, the CDT is focused on forging academic-industry partnerships, where PhD researchers conduct research which is beneficial for their partner organisation. The partnership is a requirement of a CDT-funded studentship.

The PhD's short brief outlined that government departments want to make use of algorithmic technologies due to increasing pressure on the departments' budgets, that researchers have noted these technologies can discriminate against people in intentional or unintentional and opaque ways, and that this is something government departments, including the DWP, wish to avoid. Given that their algorithmic systems are designed for the most vulnerable in society, the DWP were particularly concerned about this issue. The brief also identified the need to investigate and identify approaches to mitigating the risk of algorithmic bias, and the need for more knowledge about how to create alternative approaches and 'fairer' systems was needed. The full brief can be seen below:

> Data-driven technologies are transforming society, as governments, businesses and other sectors are increasingly adopting automated and algorithmic systems in the search for greater efficiency in the delivery of their services. Among these actors government departments, often resource-poor and in need of effective, streamlined, automated systems, are increasingly turning to digital technologies. But data-driven and algorithmic systems are far from straightforward. As a number of researchers have noted, they can discriminate in opaque ways through bias written into the systems, which can be intentional or unintentional. This is something that government departments providing support and services to the most vulnerable in society wish to avoid, but how to do so is in need of investigation. Similarly, more knowledge is needed about the expectations of citizens and related questions of ethics and trust. Using a combination of methods, this PhD project involves working closely with one such government department to explore algorithmic bias, its risks and consequences, alternative approaches, communicating about algorithmic processes with service users, and integrating alternative, or 'fairer', processes into existing workflows. The partner on this PhD project is the Department for Work and Pensions (DWP), which is responsible for welfare, pensions and child maintenance policy.

I used this brief as a basis to develop my PhD into three interconnected research papers. The short brief, and consequently overarching research aim for the PhD, was to explore how algorithmic bias can be mitigated within public sector services, and what can be learned about

this in the DWP context. In fig. A. below, I outline the research questions for each of the papers. The structure for these papers was as follows:

- *paper one*: investigate what DWP data scientists are currently doing in areas related to algorithmic bias
- *paper two*: investigate how other organisations are successfully attempting to mitigate algorithmic bias
- *paper three*: investigate how the insights from paper two could be integrated into a DWP context

The DWP's support of the project rested on their interest in research which would be useful in attempting to mitigating the effects of algorithmic bias, and the CDT funding strongly encouraged productive collaborations with industry partners. Thus, the research project would need to focus on research questions which were based in practical aspects of the DWP context, in addition to concentrating on the potential material repercussions of the development of algorithmic bias in 'real-world' usage of algorithmic technologies. Therefore, it was important that my research approached the topic of algorithmic bias from an applied perspective and engaged DWP data practitioners in the context of their everyday working lives. Furthermore, my research would need not only to address the issue of how algorithmic bias develops, but also to consider how DWP working practices might be influenced to address this issue. Considering these requirements, my overarching thesis and each research project were developed using an applied research framework (for further discussion of applied research, see Chapter 3).

Another requirement stipulated by the CDT for Data Analytics and Society funding was that the PhD would use a three-paper model approach to prepare PhD researchers with the skills necessary for academic careers, in addition to providing partner organisations with tangible research insights prior to the completion of the PhD programme. This meant I had to organise the thesis (researching how the risks of algorithmic bias might be mitigated at DWP) into three distinct research papers (for further discussion of how this was approached, see Chapter 3).

More information about the development of the brief into research questions can be found in Chapter 3 (Methodology), and more information about the 3-paper model can be found in section 3.2 (Discussion of the 3-paper model).

#	Project title	Research Questions	Methods
1	Investigating the role of current DWP working practices in mitigating algorithmic bias	RQ1a: what algorithmic bias working practices are currently engaged in by data science practitioners at the DWP?	Interviews, document analysis

		RQ1b: What are the limitations of these practices?	
2	Lessons in mitigating bias from the field: Exploring good practice and moral challenges on the AuroraAI project	RQ2a: What might 'good practice' on an algorithmic project look like? RQ2b: What challenges does good practice on an 'ethical AI' project such as AuroraAI face in practice?	Interviews, document analysis
3	The influence of DWP organisational culture on the adoption of algorithmic bias mitigation practices and implications for practice	RQ3a: What aspects of DWP organisational culture might influence the adoption of mitigation approaches? RQ3b: And, what does this mean for what might work to mitigate algorithmic bias in practice at the DWP?	Workshops, document analysis, and interviews

Fig. A

1.3. Outline of thesis

My thesis consists of seven chapters. This, the first chapter, introduces my thesis, providing a brief background to the problem of algorithmic bias in the public sector, and discussing the aims and objectives of my thesis.

The second chapter provides a literature review relating to the research topic. This chapter begins by providing a broad overview of 'algorithms' and the history of their development. After this, I discuss the concept of 'algorithmic bias' itself and how this relates to the concept of discrimination. This is followed by a discussion of how algorithms have been used within the public sector. Lastly, I discuss recent algorithmic bias mitigation efforts. In this section, I first discuss algorithmic bias mitigation methods which focus on technical approaches and identify the challenges in these approaches. Following from this, I use the FATE (fairness, accountability, transparency and ethics) framework to discuss critiques from critical data scholars and algorithmic justice activists. Throughout, I discuss how these critiques can be linked to the broader critiques of categorisation and power, within which algorithmic decision-making plays a part.

In the third chapter, I provide an in-depth discussion of the methodological approach used in my thesis. This starts with discussing the 3-paper model. Afterwards, I discuss my research approach. I then discuss my research design and data collection methods: interviews, document analysis, and workshops. In these sections I discuss the strengths of these data collection methods, why they are used in my thesis, and my sampling criteria and related decisions for each method. I subsequently discuss how I approached analysing the papers through thematic analysis. To conclude this chapter, I discuss the ethical considerations and limitations of my research.

The fourth chapter presents the first empirical paper, *Investigating the role of current DWP working practices in mitigating algorithmic bias*, focusing on both RQ1a: what algorithmic bias working practices are currently practiced by data science practitioners? and RQ1b: What are the limitations of these practices? The paper starts with a review of guidance on mitigating bias in algorithmic systems, and practical algorithmic ethical guidance more generally, followed by a description of my methodological approach to this paper. For this paper, I conducted eight interviews with DWP practitioners about two recent data projects on which my participants had worked; the Digital Trialling Framework and the Digital Plus Trial, to explore where bias might emerge and the scope that existed to mitigate it. This study identified two key findings. Firstly, my participants strongly relied on legal frameworks to guide their ethical conduct, including bias mitigation, due to their position as civil servants. However, the legal frameworks to which they were required to adhere did not facilitate accountability to the population they served. Secondly, my participants' working practices in relation to bias checking were limited by previous research conducted by the DWP, and by influences from the department's organisational culture.

The fifth chapter presents the second paper, *Lessons in mitigating bias from the field: Exploring good practice and moral challenges on the AuroraAI project*, focusing on RQ2a: What might 'good practice' on an algorithmic project look like? and RQ2b: What challenges does good practice on an 'ethical AI' project such as Aurora AI face in practice? AuroraAI is a recommender system in development by the Finnish Ministry of Finance, designed to recommend services to citizens depending on their individual circumstances. In the literature review, I identify the way that different projects have demonstrated designing algorithms is complicated, and that defining what is fair, ethical, or just is difficult. To investigate how public sector organisations are addressing the challenge of designing 'ethical' AI, I collected qualitative data through semi-structured interviews and utilised document analysis to better understand how stakeholders on the AuroraAI project led by the Finnish Ministry of Finance are responding to the challenges posed by algorithmic bias. Additionally, I interviewed AI Ethics experts, predominantly from algorithmic justice organisations, about the AuroraAI teams' proposed algorithmic bias mitigation plans.

In the paper, I identify two key findings. First, even in this purportedly progressive project, there is a lot of disagreement about what constitutes good practice in mitigating algorithmic bias and the types of solutions that might be practically implementable. These differences in understanding, combined with systemic issues such as limited funding and organisational working practices, meant the AuroraAI ethics committee felt moral concerns were sidelined.

Additionally, AI Expert participants perceived good practice regarding algorithmic bias differently. Some participants placed a stronger emphasis on consideration of the wider ecosystem of inequality than other participants. Some participants were more interested in the use of AIAs (Algorithmic Impact Assessments) as an instrument in mitigating algorithmic bias than others. The second key finding is that project management styles which focus on technological pursuits may not give enough time to focus on how to mitigate the impact of biases. These early findings move beyond existing understandings of algorithmic bias mitigation practices, which focus on either individual constraints or macro-level analysis, to highlight the importance of contextual organistional and structural constraints in public sector algorithmic bias mitigation.

The sixth chapter presents the third paper, *The influence of DWP organisational culture on the adoption of algorithmic bias mitigation practices and implications for practice*, focusing on RQ3a: What aspects of DWP organisational culture might influence the adoption of mitigation approaches? and RQ3b: and, what does this mean for what might work to mitigate algorithmic bias in practice? The literature review in this paper discusses algorithmic bias mitigation efforts such as Value Sensitive Design (VSD) and Algorithmic Impact Assessments (AIAs). In the chapter, I explore the literature on organisational change and how this is relevant to the adoption of algorithmic bias mitigation methods. The methods for this paper involved conducting a series of seven educational workshops on algorithmic bias mitigation, and seven follow up interviews with practitioners in the Department of Work and Pensions (DWP). The workshops focused on how algorithmic bias might develop, and explored the bias mitigation tools discussed in the literature review, such as algorithmic impact assessments and value sensitive design. After these workshops participants were invited to take part in a follow up interview, to allow them to reflect on the content of the workshops and its relevance to their working practices.

This third paper identifies three key findings, presented as two challenges and one opportunity. The first challenge is that it is difficult for civil service practitioners to align technologies to social justice values when servicing a large diverse public. Participants explored this issue by talking about the rights and perceived expectations of taxpayers, explaining how this group of people often had diverse and conflicting views. Furthermore, civil service practitioners' scope for action is limited by the political structures they work within, and government policy approaches may sometimes be in opposition to social justice values. The second challenge is that practitioners perceived there to be a lack of clarity within organisational guidance. They felt that anti-discrimination legislation can lead to additional uncertainty as to how conflicting needs within the population should be addressed. The opportunity identified in this paper is that participants perceived diversity in the workforce as important to algorithmic bias mitigation efforts. However, due to influences at the organisational level, some participants were uncertain as to how effective this might be.

The seventh chapter is the concluding chapter. This chapter discusses the findings suggested by each of my three empirical papers; three general conclusions to existing critical data and algorithm research have been drawn from analysing these papers alongside each other. The first is that the type of fast paced working practices found in the development of algorithmic

technologies is not conducive to the type of slower paced thinking needed to consider algorithmic bias using a socio-technical lens. Often, practitioners are under pressure to produce results quickly, and this may lead to the prioritisation of more immediately tangible results such as the project's technical deliverables. The second overarching contribution is the significance of the UK civil service context and the unique challenges which exist therein. Specifically, algorithmic technologies deployed within a civil service context are strongly influenced by political processes and build on policy decisions already put in place by government officials. Additionally, these practitioners may be required to balance the views of the public in a way private organisations do not. The final overarching contribution is that the views of the public are currently missing from discussions on how the public sector should engage with algorithmic technologies, leaving practitioners to imagine what the publics' views might be.

Following from the discussion of the overarching conclusions to my thesis, I discuss the recommendations for practice which emerged from my research. These include; 1) practitioners from different disciplines and roles need to create a shared understanding of algorithmic bias, 2) UK civil servants seeking to further social justice aims must scope out potential routes that are possible within the constraints of the civil service, 3) practitioners should seek to adopt socio-technical algorithmic bias mitigation methods, such as AIAs, VSD, and critical thinking about data and its wider environment. I then discuss the limitations of this research, and potential avenues for further research.

2.1. Introduction

This chapter presents a literature review focusing on the use of algorithmic technologies for decision making, in addition to discussing attempts to mitigate the risks of algorithmic bias inherent in these technologies.

Chapter 2: Literature review

Firstly, I provide a broad overview of 'algorithms' and their history (section 2.2). After this, I discuss the concept of 'algorithmic bias' (section 2.3), and how this relates to the concept of discrimination (section 2.4). This is followed by a discussion of how algorithms have been used within the public sector (section 2.5). Lastly, I discuss recent algorithmic bias mitigation efforts (section 2.6). In this section, I first discuss algorithmic bias mitigation methods which focus on technical approaches, and identify the challenges in these approaches (section 2.6.1). Following from this, I use the FATE (fairness, accountability, transparency and ethics) framework to discuss critiques and approaches argued by critical data scholars and algorithmic justice activists (section 2.6.2). Throughout, I will discuss how this ties in with the broader critiques of categorisation, power, and trends within which algorithmic decision-making plays a part.

2.2. What is an algorithm?

The term 'algorithm' emerges as one of the most recent buzzwords of the 21st century (Agile CRM, 2020). It has taken on myriad meanings, depending on the speaker and the context of its use. Even in 'technical' fields such as statistics, data science, and AI development, a sense of definitional blur has developed regarding the concept of an algorithm (Seaver, 2017). At its most simplistic level, an algorithm can be defined as a "description of the method by which a task is to be accomplished" (Goffey, 2008, p15). This definition is wide enough that it covers the range of processes described under the 'algorithm' umbrella; however, it provides little insight into the processes themselves. Typically, "the description" has a mathematical character – a series of formalised rules and operations to be performed to provide a desired output. The output might be a list of popular Tweets, numbers representing the likelihood of a person defaulting on their mortgage, or a selection of products associated with a buyer's previous purchases. Prior to expanding on the conceptual range regarding the term 'algorithm,' I present a brief history of the development of modern-day algorithmic technologies.

Whilst the algorithm has become a prevailing phenomenon of the 21st century, algorithms have long been linked to the automation of tasks. An early 19th century example can be found in the textile industry, when Joseph Jacquard developed a punch-card system which allowed automatic looms to output detailed fabric – a feat previously only achievable by human hand (Aikat, 2001). The punch-card acted much like a fabric pattern, providing the loom with a series of instructions it would be able to compute and follow. While initially, the impact from this development was limited to the textile industry, it later inspired computing pioneers Charles Babbage and Ada Lovelace, inventors of the modern computer program (University

of Liverpool, n. d.). This highlights how the automation of processes has historically required the creation of formalised rules based on human knowledge and actions. For the loom, this was achieved through the formalisation of the fabric pattern previously used by humans operating a manual loom. Modern-day algorithmic technologies differ in this respect; instead of formalising a simple set of instructions, these technologies formalise processes using probabilistic operations. These operations incorporate inductive reasoning and probability theory to assess the likelihood of an event occurring to approximate human decision making.

The probabilistic turn in automation occurred during the development of expert systems. These systems were early attempts in AI development, which aimed to emulate the decision making processes of human experts (Hayes-Roth, Waterman and Lenat, 1983). These systems investigated the potential for combining the knowledge of human experts (called the knowledge base) with formalised rules on reasoning (called an inference engine) to solve complex problems (*ibid*). Most expert systems used a process whereby the knowledge base was constructed through interviewing domain experts, followed by the formalisation of this knowledge through the construction of mathematical models based on the experts' problem-solving processes (*ibid*). However, a small number of systems experimented with automating the process of knowledge acquisition using Machine Learning (ML). Machine Learning approaches could construct the knowledge base for an expert system using inductive probability theory with large specialist datasets. These automated knowledge acquisition systems were successfully used to identify new compounds in chemistry, in addition to new approaches to designing computer chips (Buchanan *et al.*, 1976; Hayes-Roth, Waterman and Lenat, 1983). These tentative steps towards integrating probability theory into expert systems would lead to the development of modern-day Machine Learning (ML) approaches.

To understand the mechanics behind modern algorithmic technologies further, it is necessary to explore how machine learning algorithms work. Machine Learning has three specialist branches: supervised, unsupervised, and reinforcement learning. In the context of this thesis, I focus only on supervised and unsupervised machine learning approaches. A typically used example to explain the mechanisms of supervised machine learning can be found in algorithms which filter emails into 'spam' and 'not spam' (Kelleher and Tierney, 2018). In this example, a supervised machine learning algorithm will be given a dataset of both spam and not spam emails. This is referred to as the 'training' dataset, as it is used to 'train' the machine learning algorithm. In this dataset, each email will be given a label which describes it as spam or not spam, which will be the 'target value'. The supervised machine learning algorithm then attempts to find the most appropriate mathematical function[1] which maps the attributes (i.e. images, hyperlinks, address endings) of those emails onto the emails' target value (spam/not spam). Or, to put it another way, "the function the algorithm learns is the spam-filter model returned by the algorithm" (Kelleher and Tierney, 2018, p99). In other words, the supervised

[1]A mathematical function is a formalised description of the relationship between an input and output variable (expressed as Y=f(X)). For example, the relationship between the input variable (X) and output variable (Y) might be X+2. This would be expressed as f(X) = X+2, the 'f' in this circumstance meaning function. The machine learning algorithm is given the variables X and Y, and instructed to find the relationship between these variables. The output of this is called a function.

machine learning algorithm will create a series of mathematical instructions to be used in a spam filter program based on the training dataset it was given. This spam filtering function is sometimes referred to as a model. Following from this, the accuracy of this function will be tested against a segment of the dataset which was not included in the 'training' dataset. This previously unused segment of the dataset is referred to as the 'test' dataset. If the spam filtering function learnt during the 'training' phase performs well on the 'test' dataset, meaning it has a high number of correctly identified emails, then the spam filter can be used 'in the wild'. This means it might be considered suitable to be used outside of the training data that was used to create it, in a 'real world' context.

In unsupervised learning the training dataset is not assigned a target value, and instead the algorithm searches the input data for patterns. One of the most used examples of these algorithms is called cluster analysis, "where the algorithm looks for clusters of instances that are more similar to each other than they are to other instances in the data" (Kelleher and Tierney, 2018, p102). For example, the algorithm might analyse a dataset of music listening habits, and then create categories of similar types of music listeners. The key difference between these algorithms is that the data scientist does not create the categories prior to using the algorithm, but rather the categories come from the algorithms' assessment of similarity between instances in the dataset. For example, the algorithm might find a relationship between users being age 30 – 40 and listening to nu-metal. An example of how this technique has been used in the public sector can be found in law enforcement, where an unsupervised machine learning algorithm might be used to search for patterns in crime data and provide a series of categories which provide information as to where and when crime may occur.

During the 1980s-2000s, statisticians and computer scientists debated the value of these types of 'algorithmic models' against the more traditional approach of 'data modelling' (Breiman, 2001). In his paper *The Two Cultures*, Breiman (2001) described 'data models' as models which are manually created by a statistician, instead of an algorithm, where a statistician uses statistical techniques to infer the relationship between variables of a known phenomenon. He characterised data modelling methods as ones which focused primarily on theory, and the work of statisticians as primarily choosing the variables included in their models based on what is theoretically known about the subject (*ibid*). However, Breiman (2001) believed this process led to the creation of models which had very little practical use outside of developing theory. He compared this to algorithmic modelling methods, such as machine learning, which in his opinion had far greater practical value, due to his perception that they had far greater predictive power and could be used to predict future trends (Breiman, 2001). Many of these algorithmic techniques rely on methods which are less interpretable by humans, but which performed better predictively through their use of 'test' and 'training' datasets (Breiman, 2001). He argued these algorithms better replicated the "black box of nature", a phrase he used to describe how he perceived phenomenon actually worked in the natural world (*ibid*). This contrasts with how he perceived the work of statisticians, who attempted to replicate "nature" by picking a very limited number of variables they thought were most important in the phenomenon they studied. Thus, he

thought algorithmic methods were more successful than data models regarding inference and prediction, due to the unknowable complexity of their workings, which in his view mimicked nature itself.

The academic responses to this paper are varied. Some suggest he mischaracterises the process of creating data models, others assert the need for a plurality of quantitative processes within the scientific method, and discuss their experiences of the difference in approaches which can be found in academia compared to industry (Breiman, 2001). This disagreement surrounding the place of algorithmic modelling techniques can be described as an epistemological disagreement on the processes used to ensure the validity of quantitative scientific practice. While most academics and practitioners would acknowledge the complexity within this debate, two crude camps of thought had emerged: one that believed theory could be developed based on how well a model predicted the future, and another that believed it was more important to analyse past results. Furthermore, this disagreement centred on the practical use of the theoretical understanding developed when using data models outside of scientific debate. For example, in the paper Breiman questions whether doctors gain more by understanding how a model works, or by how accurately they predict patient survival rates. Additionally, this disagreement focused on the importance of how well an analyst could explain the results of their model (Breiman, 2001; Pietsch, 2016; Hooker and Mentch, 2021). This debate on explainability is relevant to my research, and thus will be further discussed in section 2.1.2 (FATE: Transparency).

The debate on the accuracy and epistemological significance of machine learning algorithms has long raged within the fields of computer science and statistics. However, these methods have also been of great interest to academics in more sociological fields. In Gillespie's (2014) paper *The Relevance of Algorithms,* he identifies why algorithms are relevant objects for sociological study (Gillespie, 2014). The most salient in the context of this thesis are; patterns of inclusion, the promise of algorithmic objectivity, and entanglement with practice. I discuss these in turn below.

The first of these, *patterns of inclusion*, focuses on how the developers of algorithmic technologies make decisions about what data are or are not included within an algorithm's design (*ibid*). Gillespie (2014) describes how despite the growth in available data in the 21st century, data must still be collected, selected, and processed prior to being analysed by an algorithm. He argues that this process is not neutrally carried out by machines, but instead relies on human judgement about what to include or exclude as part of the data collection and pre-processing process. This argument is also made by D'Ignazio & Klein (2020) in *Data Feminism*, where they discuss how structural biases are reinforced by the types of data organisations collect. In a chapter entitled "What gets counted counts", they describe how data collection practices reinforce cultural norms. For example, when organisations use data collection instruments which only provide the option of two genders (typically female or male), this reinforces the importance of the gender binary, and excludes non-binary identities from public recognition.

The second of Gillespie's (2014) factors I discuss is *the promise of algorithmic objectivity,* which focuses on how algorithms are portrayed as neutral and objective. He argues that algorithm providers positioned their algorithms as objective decision makers and their evaluations are portrayed as "fair and accurate, and free from subjectivity, error, or attempted influence" (Gillespie, 2014, p179). This positioning, he argues, lends algorithmically produced results a sense of authority and legitimacy with respect to their accuracy, and masks the political character of the processes and decisions which are embedded within algorithmic practice. This follows a long history of portraying technologies as objective and neutral. In 1988, Donna Haraway (1988) described how technology is built from a particular standpoint and is designed to enhance a particular way of seeing the world:

> "Histories of science may be powerfully told as histories of the technologies. These technologies are ways of life, social orders, practices of visualization. Technologies are skilled practices. How to see? Where to see from? What limits to vision? What to see for? Whom to see with? Who gets to have more than one point of view? [...] Struggles over what will count as rational accounts of the world are struggles over how to see" (Haraway, 1988, p587)

Developments in algorithmic technologies typify the struggle over what counts as a rational way of seeing the world. As discussed earlier, some computer scientists strive for their approaches to mimic the "black box of nature" (Breiman, 2001). They assert that algorithmic technologies imitate natural processes. However, as described by Haraway (1988) above, this masks the viewpoints embedded within these technologies. For example, to return to D'Ignazio & Klein's (2020) argument that "what gets counted counts", when data scientists use data which reinforce the gender binary, this privileges and rationalises this way of seeing. Moreover, it positions this way of seeing as the natural order of the world.

Additionally, organisations' appeal to objective authority regarding evaluations produced by algorithms masks the actual processes, procedures, and structure of modern-day algorithms. The early algorithms Breiman (2001) describes are, for the most part, self-contained. They perform a particular task, in conjunction with other algorithms or teams of people. This is not true of all algorithmic technologies, with some modern day "algorithms" spanning entire organisations. In Seaver's (2017) paper on algorithmic practices at Spotify, he describes how among the data workers he interviewed, none believed they were the ones working on the 'algorithm'. He argues that due to the scale at which these algorithms are now constructed, it is difficult for those working on the algorithm to take possession of their work on the said algorithm. He proposes that the modern-day algorithm is not simply a model and an adjoining dataset, but rather a collection of teams, cultural working practices, and technologies.

The third of Gillespie's (2014) factors I discuss is *entanglement with practice*, which focuses on how users might change their behaviours to suit, or resist, the algorithms with which they engage. He argues algorithms cannot be understood as merely static processes, which have a one way "effect" on users who passively receive the judgements of algorithmic processes within the context of their lives (*ibid*). Rather, he argues that it is important to recognise the "entanglement" that exists between users and algorithmic technologies (*ibid*). Users change

their behaviours and how they interact with an algorithm depending on their own needs, wants, and motivations, in addition to their expectations of the algorithmic technology in question. In a qualitative study (2017), Bucher found Facebook users' timeline algorithms would, on occasion, 'act' in a way which had a direct influence on the users' mood, for example, by reminding them of an ex-partner. This awareness of the algorithm would sometimes change how users interacted with the platform (Bucher, 2017). This demonstrates a complex feedback loop between humans and algorithmic processes; as these algorithms learn from users' behaviour, users can become aware and alter their behaviour in response, which then has an effect on the algorithm as it reacts to new patterns.

In conclusion, the term algorithm has been the subject of a wide range of debates. Within the data science community, these have included questions about the validity, epistemological assumptions, and best working practices when utilising algorithmic methods. Within the social sciences, academics have questioned the supposed objectivity of algorithmic methods, in addition to how they are embedded within social structures. Furthermore, this section has highlighted the wide range of processes which may be referred to as an 'algorithm'. Within the context of my thesis, I primarily use the word algorithm to mean a mathematical model which is either used in automated decision-making practices, or a data model which is used at scale to make decisions regarding individuals (e.g. the UK Ofqual A-Level case). Furthermore, I refer to an algorithm as a process which may perform a wide range of tasks, including what may appear in one's recommended tweets page through to calculating risk scores of various types. I use the term 'algorithmic technologies' to describe a technology which contains an algorithm.

In the following section, I explore the concept of algorithmic bias, through considering how algorithms perpetuate bias, in addition to how the phenomenon of algorithmic bias relates to the concept of discrimination.

2.3. What is algorithmic bias?

As algorithmic technologies increasingly shape our lives, prominent AI policy organisations such as AI Now and Algorithm Watch warn of a phenomenon known as algorithmic bias (Campolo *et al.*, 2017; Jaume-Palasí and Spielkamp, 2017). Algorithmic bias describes a situation where a model consistently discriminates against a group of users. For example, in ProPublica's 2015 investigation into the use of algorithmically calculated risk assessments in the American criminal justice system, algorithms were found to discriminate against defendants on the basis of race. These algorithmically calculated risk assessments were used to inform judges as to the defendants' likelihood of reoffending (their recidivism risk), to inform a judge's decision-making process when setting the defendants' bail bond amount. ProPublica found the recidivism risk algorithm used consistently rated black defendants as more likely to reoffend than white defendants – even when defendants had similar criminal histories (Angwin *et al.*, 2016: Kirkpatrick, 2016). The white defendant would be given a lower or equivalent reoffending score than the black defendant, even when the black defendant

had a less extensive criminal record (Kirkpatrick, 2016). This had the effect of consistently and systematically discriminating against black defendants.

Concerns about algorithmic bias are not new. Automated decision-making systems discriminating in this manner were found as early as 1986 in a hospital computer program designed to select interview applicants based on statistical methods and past applicant success rates (Asscher, 1988). This programme was found to systematically discriminate against black and female applicants. This discrimination was caused by the lack of applicants with these characteristics who currently worked within the hospital, leading to a subsequent lack of available training data including individuals with those characteristics (*ibid*). Furthermore, Friedman & Nissenbaum (1996) discuss the issue of bias in computer systems as early as 1996. In their paper *Bias In Computing Systems*, they put forward a taxonomy of three types of bias: pre-existing bias, technical bias, and emergent bias. Pre-existing biases are the type which exist within a structure or organisation, and become embedded within the computing system. Technical biases are those imposed by the technical limitations of the practitioners working on the project. Emergent bias is the type of bias which develops from the system being used in a new context (Friedman and Nissenbaum, 1996).

Algorithmic technologies can also perpetuate inequalities in more subtle ways than the previously discussed US criminal justice example. Google on average handles 3.5 billion searches per day (Internet Live Stats, 2019), and is a prominent source of information to large parts of the world. In *Algorithms of oppression: How search engines reinforce racism*, Noble (2018) examines how search engines prioritise results which contain biased assumptions regarding race and gender. For example, she found when she searched the phrase 'black girls' using Google search, the PageRank algorithm was highly skewed towards sexualised and pornographic content, placing this type of content at the top of the results page (p17). In her work she uses a Black feminist lens to explore how white supremacy and sexist culture perpetrates itself throughout the internet, despite prevailing discourses of the internet as a fair and equal marketplace of ideas (p84).

This type of representational bias has been explored by Otterbacher *et al*. (2017), who found that normative gender biases were reproduced in Bing search results. In their study, they found results for warm-traits (such as "emotional," "expressive" and "sensitive") were more likely to retrieve images of women. In contrast, competency-based traits (such as "ambitious," "intelligent" and "rational") were more likely to retrieve images of men (Otterbacher, Bates and Clough, 2017). They state that it is unlikely these search biases come from the algorithms themselves – but instead are an extension of the social and historical structures they are embedded within (Otterbacher *et al*., 2017, p9: Noble, 2018, p13). As discussed earlier in relation to Bucher's (2017) research on algorithmic technologies, the way in which users interact with the algorithm affects how algorithms behave, and can cause what she refers to as 'feedback loops.' In the Google search examples, users clicking for content which reflects their own social biases can lead to this content appearing prominently within the search engines' results. As these results are prominent, it may then lead to more users clicking on these links, thus creating a 'feedback loop.'

Despite examples of what algorithmic bias *looks like*, there has been difficulty pinning down the cause of algorithmic biases. Indeed, early comments from Danks and London (2017) on the subject included that "[p]ublic discussions of algorithmic bias currently conflate many different types, sources, and impacts of biases, with the net result that the term has little coherent content" (Danks and London, 2017, p2). They outline some common causes of what is known as algorithmic bias, including: bias caused by biased training data; bias caused by inputting protected characteristics; biases caused by inappropriate deployment of an algorithm; and bias caused by misinterpretation of the algorithm's output (Danks and London, 2017, p4). However, they also state that algorithm bias is impacted by statistical, ethical, and legal biases. Furthermore, they state that due to the lack of coherent notation of 'algorithmic bias,' "there is little reason to think that there is one consistent or reliable response to these myriad possible biases." (Danks and London, 2017, p2). They also propose 'de-biasing techniques' – that is, using intentionally statistically biased training data or techniques to help counteract historical bias when deployed. I return to discussing these types of de-biasing techniques in section 2.1.1 (Technical Mitigation Methods).

While the above discussions are useful for understanding the different types of bias from a technical standpoint, AI Now's 2017 report highlights "[t]he word "bias" also has normative meanings in both colloquial and legal language, where it refers to judgements based on preconceived notions or prejudices" (Campolo et al., 2017, p14). Focusing solely on a technical understanding of 'bias' ignores the prejudices or assumptions of designers that can be "wittingly or not, frozen into the code, effectively institutionalising those values" (O'Neil, 2017). As touched on earlier, data scientists often strive to maintain a 'neutral' position in their work, seeing themselves as merely engineers or seeking to take an apolitical stance to their work (Green, 2018). However, Green (2018) argues that if data scientists do not confront the social assumptions embedded within these technologies, and instead only focus on bias as a technical issue, issues of algorithmic bias cannot be addressed: "striving to be neutral is not itself a politically neutral position—it is a fundamentally conservative one." (Green, 2018).

To complicate the issue further, it has been difficult to know how algorithmic bias happens, due to the black boxed character of algorithmic technologies. By this I mean that it is difficult (or impossible) to know what processes contributed towards the biased outcome. This complicates the issue of defining the technical processes which have led to algorithmic bias within a particular algorithm for two reasons. One, the algorithmic technology might rely on an algorithmic technique which is mathematically complex and hard for humans to interpret, and harder still for humans without the requisite background in that specific algorithm to understand an explanation of how it came to its decision (Wachter, Mittelstadt and Russell, 2017). As discussed earlier, Breiman (2001) argued that this complexity, and the resulting lack of interpretability, was the reason why algorithmic technologies were more accurate in their predictions than data models. While the discussants to his paper questioned the validity of this argument, it is worth remembering there is a culture of thinking within the field of data science which associates complexity, and thus unknowability, with improved accuracy. Two, the design of the algorithmic technology might be an industrial secret, and so organisations

may be reluctant to explain the algorithm's inner workings to those outside of the company (Veale, Van Kleek and Binns, 2018).

In view of these difficulties, I take on a wide definition of the concept of 'algorithmic bias.' I use the term to mean discriminatory or unjust outcomes caused by the automation of statistical techniques which aim to predict outcomes based on historical data and probabilistic processes. While some interpretations of algorithmic bias limit the scope of this concept to machine learning algorithms, this would exclude many of the older instances of data models or computer programmes producing similar outcomes, which I choose to include in my definition.

Furthermore, the wide scope of technologies included in this definition have been chosen because of two additional conceptual difficulties. The first of these is that there is a wide range of statistical techniques and processes which fall within the algorithmic technology umbrella. Some of these include simple algorithms, such as linear regression algorithms, which are similar to those worked on by human statisticians or data scientists. Others, like neural networks, are far more complex, and the results of these algorithms cannot intuitively be understood by most human beings. Secondly, the type of algorithm which has been used in an algorithmic technology is not often known, as the inner workings of these algorithmic technologies are often trade secrets. Due to the range of techniques which could be utilised in any algorithmic technology, it is unsuitable to narrow the field of inquiry to machine learning algorithms specifically. This is particularly true because it can be difficult to assess whether machine learning has been used at all, or whether the organisation has simply used a data model instead.

However, the wide range of possible algorithms or statistical techniques used in algorithmic technologies have two important factors in common. One, the decision-making models they produce are based on datasets, statistical processes, and probability theory. Two, the statistical patterns which are identified in these datasets are then formalised into a model which will be used for decision-making. This model will use the patterns identified in the collected data, to determine decisions which impact people on an individual level. In other words, statistical averages will determine the outcome of decisions made about individuals.

Given my definition of algorithmic bias, it is important to unpack the term 'discrimination'. I address this in the following section.

2.4. Discrimination

One of the thornier parts to the problem of algorithmic bias is that often the very purpose of algorithmic technologies is to discriminate (i.e. differentiate) between different groups. This can be seen in some of the examples discussed earlier, such as the applicant filtering software, and search filtering algorithms. In practice, this might also involve classifying individuals in a way which endeavours to calculate their 'risk' level- be it risk of fraud, child neglect, or reoffending; making decisions that could impact on various realms of life from financial to familial.

In *Automating Inequality*, Eubanks (2018) argues that algorithmic technologies create a 'digital poorhouse', where those who are subject to their decisions are often punished by the system. Eubanks (2018) contends that this 'digital poorhouse' builds on earlier innovations in poverty management, such as physical poorhouses, with the purpose of reinforcing a narrative of their being an 'undeserving' and 'deserving' poor. In other words, these technologies reinforce the idea that not all living in poverty deserve assistance. For example, an algorithm was developed to assist in housing the most vulnerable in Skid Row, Los Angeles[2] (Eubanks, 2018). This algorithm was developed to assess who was particularly vulnerable, giving people a score between 1 and 17 which represented their vulnerability to death or admittance to a hospital. Homeless people who wanted to apply for social housing would need to complete a questionnaire, which collected a wide range of personal data, to be put on a waiting list for the scant amount of housing which became available. However, due to the system, some applicants never receive housing, as the system judges them to be the least at risk of harm whilst homeless. By utilising this system, state officials are able to claim that the most vulnerable are housed – those most 'deserving' of housing – yet this obscures the fact only a very small percentage of homeless people living in Skid Row ever receive housing.

Prior to the development of the social safety net and national insurance schemes, the provision of financial support in case of death or misfortune lay in the hands of private organisations. These private insurance companies needed means to assess the likely lifespan of an individual, to balance premiums against payouts and ensure the profits of their shareholders (Wilson, 2018). In *'Babbage among the insurers: Big 19th-century data and the public interest'* Wilson (2018) posits that the data compiled by life assurance companies to provide accurate and reliable information on the average lifespans of a given group of people were a forerunner to the large personal datasets of today (Wilson, 2018). However, in Victorian England, the data collected by life insurance companies was still limited, and originally these companies relied on simple mortality data combined with the Law of Large Numbers[3] to predict the collective lifespan of those they insured (Alborn, 2009). As these methods developed, insurance companies later hired physicians to perform medical exams. This provided them with further data such as height, weight, and any underlying conditions – allowing them to classify individuals within a broad range of categories relating to the individual's health (or decided lack thereof) (Alborn, 2009).

Like the narrative of the deserving and undeserving poor that Eubanks (2018) notes, this created a divide between the insurable and uninsurable. Although this application of the Law of Large Numbers was deemed 'enlightened' when originally put into use (Wilson, 2018), it served to decide who was too risky to insure and those who had to pay higher premiums. Indeed, eugenicists and social hygienists called upon life insurance companies to share their information in pursuit of "shifting the mean towards the tail that approached perfection" (Alborn, 2009, p302). To put it another way, eugenicists wanted this data for the purposes of statistically analysing human health and reducing the number of people whose health they

[2] A 62 block area in LA which houses over 8000 homeless people, primarily in open tents (Cristi, 2019)
[3] The law of large numbers, in probability and statistics, states that as a sample size grows, its mean gets closer to the average of the whole population (Investopedia, 2019).

judged to be below average. Part of the power behind these metrics is their creation of norms based on averages, providing an 'objective' veneer, that enhances their credibility (Porter, 1996, p78). This apparent objectivity also further strengthens the grip of meritocratic thinking, as it provides neoliberalism with a scoreboard and a set of seemingly fair rules, and reinforces individualist attitudes (Beer, 2016, p195: Kendell, 2015). The request of eugenicists was, however, judged to be unprofitable – and life insurance companies settled on "healthy enough" (*ibid*).

This attempt to discriminate between different types of customer – those who were judged too risky to insure, and those judged safe to insure – touches on discrimination as it is often understood today. In colloquial usage, discriminating against someone is often linked to holding a social prejudice about a particular group of people (Campolo *et al.*, 2017). For example, as seen in the history of insurance, holding a prejudicial belief that people with certain medical conditions are less valuable to society than those without those conditions.

Within the context of algorithmic bias, discrimination is often understood through a legal framing (Wachter, Mittelstadt and Russell, 2017). From a UK legal standpoint, discrimination can be understood as the prejudicial or unfavourable treatment of people on the basis of "protected characteristics" (The Equality Act, 2010). The protected characteristics stipulated in The Equality Act (2010) in the UK are gender, race, age, disability, sexuality, and in Scotland, socioeconomic background (The Equality Act, 2010).

A concept closely related to that of discrimination is social inequality. Social inequality describes when the distribution of resources, or access to resources, takes place in a way which is not equal. Often, this is on the basis of protected characteristics. Discrimination law is one tool employed to advance social equality, by making it harder to refuse people access to resources such as money, power, and visibility. When algorithmic technologies discriminate against individuals, they may also entrench social inequality on a collective level (Eubanks, 2018). Furthermore, citizens' perception of social inequality has been linked to how they interpret whether data uses are unfair. In a 2022 report from the *Living With Data* project, the authors found some of their participants perceived data uses that negatively affected or discriminated against those who were already disadvantaged as being unfair (Ditchfield *et al.*, 2022). This highlights the importance of citizens' perceptions of inequality in how they assess the fairness of data uses, and by extension algorithmic technologies. Discussion of the concept of fairness can be found in section 2.7.2 (FATE: Fairness).

This section has focused on the concept of discrimination. To better understand the risks and concerns of algorithmic bias within the public sector, in the following section I discuss how algorithms are used in the public sector.

2.5. Algorithms in Public Services

In recent years, public sector services have increased their use of algorithmic technologies. In 2018, The Data Justice Lab reported at least 53 UK councils had previously used data analytics (Dencik *et al.*, 2018). Uses included, but were not limited to, targeting police efforts, providing

clinical insights to the NHS, answering customer queries within government services, enhancing decision making in areas such as child protection and policing, and identifying troubled families (Dencik *et al.*, 2018; Oswald *et al.*, 2018; Hughes, 2019). Furthermore, in 2020, The Guardian reported that responses to their freedom of information requests indicated 100 out of 226 UK councils have used algorithmic systems to assist in their operations (Marsh and McIntyre, 2020). I focus on an example of a child protection algorithm below, because of the recent prominence of the adoption of these systems.

In 2018, Hackney Council (UK) implemented a child protection system known as the Early Help Profiling System (EHPS), which provides the council with a method of identifying children who may be at risk of abuse or neglect. It is claimed that this system provides social workers with risk scores for each family, so that help and resources can be better targeted towards those most in need (Dencik *et al.*, 2018). This system was developed with the help of Ernst & Young and Xanture (two third party data companies), and uses statistics including "information about school attendance and attainment, families' housing situations, and economic indicators, and turns them into risk profiles for individual families." Full automation is estimated to provide $160,000 worth of savings in staff costs (Apolitical Foundation, 2017).

Systems which provide risk scores for child welfare are becoming increasingly common. Eubanks' investigated similar systems used in the USA. She describes one where data from various sources (such as previous history with the service, school attendance, and so forth) are used to create a score which reflects the supposed risk of a child being neglected or abused (Eubanks, 2018, p127). The score is then passed on to a case worker, who can use it when judging whether an investigation should take place. To make these assessments, the algorithm uses proxy measurements to provide insight into the individuals' behaviour. However, these proxy measures are often a reflection of poverty itself, such as not having the money to feed or clothe children (Eubanks, 2018). Thus, the system categorises families living in poverty as being highly likely to commit child abuse or neglect merely as a result of living in the material conditions of poverty (*ibid*). Furthermore, Eubanks (2018) notes two prominent problems with the system; the score is highly influenced by the family's previous history with the service, and humans defer to these scores when unsure.

Part of the appeal of these systems is that they seem to offer the solution to distributing scarce resources in an effective and fair manner (Dencik et al., 2018, p120). However, Eubanks (2018) has contended that the digital poorhouse encouraged by algorithmic technologies instead weakens the premise of the social safety net. Eubanks argues that when algorithmic technologies are used to distribute resources, this goes against the premise the social safety net is built upon. That is, the societal agreement that the costs of uncertainty should be shared between society's members (*ibid*). By using technologies which aim to assess which citizens are most deserving of assistance, society is prevented from "[sharing the] collective responsibility for creating a system that produces winners and losers, inequity and opportunity." (Eubanks, 2018, p198).

Additionally, Parks and Humphry's (2019) study examines how exclusionary design practices within algorithms can contribute towards a new digital divide, through the intersection of

technological and social issues. One of these cases is colloquially known as 'Robo-debt', a project which automated the detection of overpayments within the Australian welfare system, shifting the processes to an online service, and automating the delivery of debt collection notices without any human oversight as to their accuracy. The new automated system caused an overwhelming amount of anxiety and stress for citizens, with little face-to-face support available to answer queries or opportunity to prove their income for the time period detected by the algorithm. Thus, citizens were not able to contest the decisions made by the algorithm. Specifically, this deepened existing inequalities as it did not factor in the impact of disability, digital literacy, and the challenges brought about by socioeconomic exclusion on ability to use digital systems into its design process. Alongside changing the automation technique which investigated overpayments, the designers changed fundamental principles of how the system worked – which included placing the burden of proof on citizens to prove their earnings, rather than the organisation needing to prove the welfare recipient did indeed owe them money (*ibid*, p12). This highlights the way that what may seem like simple techno-bureaucratic mechanisms have very tangible results in citizens' ability to engage with public services.

As members of the public rely on public services for key areas such as health, tax, and housing etc, public perception of algorithmic technologies is important. Thus far, public perception of the use of algorithmic technologies in public service contexts remains mixed, and highly dependent on the context they are employed within. Kaun *et al*. (2023) investigated public perception surrounding algorithmic technologies within a welfare context using survey data from three different countries (Germany, Estonia, and Sweden). The results of this analysis demonstrated differences in levels of trust in these systems in each country (Kaun, Larsson and Masso, 2023). Furthermore, Kaun *et al.* (2023) theorise that these differences are tied to respondents' past experiences of public sector services within their country. Similar results have been found using qualitative methods in a UK context, such as in the *Living With Data* (2023) report (Ditchfield *et al.*, 2022). In this report, some participants from a disadvantaged group or minority background were found to link their experience of discrimination to potential ill-use of their data (*ibid*). Moreover, it was found that some participants who were not from a disadvantaged group were still concerned about unfair data use, and utilised their understanding of social inequality to imagine what issues those from disadvantaged groups might face.

In the following section, I turn to discuss the methods proposed to mitigate the risks of algorithmic bias.

2.6. Algorithmic bias mitigation efforts

In this section, I discuss proposed methods to mitigate the impact of algorithmic bias. I first discuss technical mitigation methods (2.6.1). By this, I mean algorithmic bias mitigation methods which use technical skills such as dataset de-biasing. Following on from this, I explore

algorithmic bias mitigation methods which are more social in character, using the FATE (fairness, accountability, transparency and ethics) framework (2.6.2).

2.6.1 Technical mitigation methods

Thus far, many solutions to algorithmic bias have been technical in character. Technical algorithmic bias mitigation comes in three main types; de-biasing the datasets used to train the algorithm; changing the process of optimizing an algorithm; and changing the outputs of the algorithm in the deployment phase of the technology (Balayn and Gürses, 2021).

Regarding de-biasing the datasets, a few different approaches have been put forward. Some focus on removing the data regarding individuals' protected characteristics, with the intention of preventing the algorithm from inferring on the basis of these characteristics (*ibid*). Others have suggested de-biasing techniques which focus on making datasets more representative of their target population (Galhotra, Brun and Meliou, 2017). Furthermore, some data scientists have developed techniques which switch protected characteristic variables of individuals within their datasets, to make it more difficult for the algorithm to infer on the basis of these characteristics (Balayn and Gürses, 2021).

However, the effectiveness of these de-biasing techniques has been contested. In *Weapons of Math Destruction*, Cathy O'Neil (2017) argues that data such as postcodes can be used as 'proxy data'[4] for characteristics such as socioeconomic background and race, due to their being a larger presence of some demographic groups in some areas compared to others (O'Neil, 2017). To combat this, it has been suggested that variables which correlate with protected characteristics are also removed from training datasets (Balayn and Gürses, 2021). However, this solution may cause further difficulties, as without protected characteristic variables, it becomes difficult for data scientists to assess or audit their models for discriminatory effects (*ibid*).

Turning to de-biasing techniques which focus on optimizing the algorithm, some approaches have focused on finding new ways to operationalize the concept of fairness within a statistical framework, allowing practitioners to better perform statistical checks on their models (Bellamy *et al.*, 2019). In statistical parity-based fairness approaches, statistical outputs from various demographic groups are compared to assess whether their outputs are similar. These statistical outputs can include metrics such as the number of false negatives and false positives generated by different demographic groups, or the use of other metrics such as error scores representing how well the model fits the available data (Green and Hu, 2018).

It has been argued that algorithmic bias mitigation methods which utilise an algorithmic or datafied framing of algorithmic bias (i.e. technical de-biasing approaches) are ineffective at mitigating algorithmic bias (Balayn and Gürses, 2021). Selbst *et al.* (2018) describe an algorithmic or data-based framing of bias as an 'abstraction trap.' An abstraction trap can be

[4] Proxy data is where it infers about a protected characteristic from a piece of data which is not the characteristic itself.

understood as an error made when data practitioners "[abstract] away" aspects of the social context as part of the process of constructing a mathematical model of the problem they are working on (*ibid*). One of the abstraction traps described by Selbst *et al.* (2018) is the 'framing trap', which describes how data practitioners attempt to solve social issues such as 'fairness' using methods typical of their discipline, such as altering modelling choices or using de-biasing techniques with the currently available data.

The framing trap can be understood to form the basis of much of the critique surrounding how data practitioners currently focus on mitigating algorithmic bias within their working practices. In the following section, I discuss these critiques in addition to algorithmic bias mitigation methods which go beyond this technical framing.

2.6.2 Fairness, Accountability, Transparency & Ethics (FATE)

One of the movements which seeks to address concerns about algorithmic bias is the FATE movement, which focuses on understanding the challenges presented by algorithmic technologies by focusing on Fairness, Accountability, Transparency and Ethics across algorithmic systems in use. This is an interdisciplinary movement that has brought together data scientists, computer programmers, social scientists, law experts and the humanities, to examine and challenge how algorithmic systems are being used in sectors such as education, public services, and healthcare (ACM FAT*, 2019). In this section, I use the FATE framework to discuss the concepts of fairness, accountability, transparency, and ethics in turn. In each section, I explore these concepts as proposed algorithmic bias mitigation methods, and the challenges present in these methods.

2.6.3 Fairness

One of the key concepts discussed within the FATE framework is fairness – however this everyday concept has proven far from simple. While in lay speech the word 'fair' is often used to mean something which is free from bias, dishonesty, or injustice (Dictionary.com, n.d.), it has taken on specific connotations within debates on algorithmic bias. When defined and operationalised by data practitioners, who are often core designers of the systems being discussed, two prominent operational definitions of fairness have emerged; procedural fairness and statistical fairness (Green, 2018; Green and Hu, 2018). Procedural fairness is concerned "with the fairness of the steps, input data, and evaluations made in a decision-making process" (Rovatsos, Mittelstadt and Koene, 2019, p11). To put it another way, this is fairness by way of process – the procedure for handling data is deemed fair, because the procedure stays the same for all groups. This differs from statistical fairness, which relies on producing metrics, such as accuracy scores and comparative statistics, to allow for comparison between different groups, to ensure treatment of all groups has been equal (Green and Hu, 2018, p2).

Green and Hu (2018) argue that both of these methods "capture important considerations of fairness: impartiality of process on the one hand and protection from adverse impact on the

other" (Green and Hu, 2018, p2). However, both of these approaches are practiced through basic procedural methods that do not move beyond the data itself – either a method of steps performed, or a method that relies on the comparison of relevant metrics the organisation or individual has chosen to rely on. Green and Hu (2018) conclude that both of these definitions are restricted to technical understandings of fairness, and by turning away from the broader context of injustice "we run the risk of overlooking systemic issues and deeming social structures fair simply because we have improved one component of them" (Green & Hu, 2018, p4). Furthermore, they argue that these conceptions of fairness draw the focus away from the material conditions of inequality and injustice described within the dataset, allowing actors to position the problem of algorithmic bias as one caused by 'bad algorithms' or 'bad data' (Green and Hu, 2018; Hoffmann, 2019). In doing so, the importance of the beliefs, values, and social biases embedded within institutions and practitioners is minimised (*ibid*).

Hoffman (2019) argues that striving for fairer algorithmic systems does not go far enough, and that systems focusing on parity or statistical equality create a narrow field of inquiry, which limits practitioners' ability to recognise how data and algorithms connect to the wider issues of injustice within society. For example, methods which focus on achieving a similar number of false positives and negatives for all groups ignore the fact that for some groups, the process of appealing decisions is far more arduous and likely to impact these groups in a more substantial way (Hoffman, 2019; Costanza-Chock, 2018).

In *Towards Data Justice*, Dencik *et al.* (2016) argue that the issue of algorithmic bias should be understood through the framework of justice, not fairness (Dencik, Hintz and Cable, 2016). They suggest that a justice-based framework connects the issues regarding data-driven practices to those of inequality and exploitation more generally. Additionally, they suggest a justice-based framework provides a conceptual foundation for creating tools to address these issues. It does this, they argue, by providing a way of examining the ideological basis of data-driven processes and considering power relations, interests, and political agendas within the context of data-driven practices (*ibid*). From there, it provides a foundation to question how society should be organised (Dencik *et al.*, 2022; Milan & Treré 2019, 2021; Treré 2019).

In alignment with this thinking, it has been argued that a justice based framing of moral issues provides a way of thinking about data which allows for 'de-centering' data (Peña Gangadharan and Niklas, 2019). Gangadharan and Niklas (2019) argue that de-centering data allows for greater recognition that discrimination, unfairness, and injustice are not primarily felt through the medium of algorithmic technologies (*ibid*). For example, within their study of civil society representatives' perceptions of de-centering technology regarding discrimination, they found that although there is concern that algorithmic technologies can be used to "[deny] benefits to as many people as possible (p2)", interviewees were less concerned about the technology itself, but more with the already existing problems within the public administration (*ibid*). This echoes the findings of the *Living With Data* project, a qualitative research project which investigated public perceptions of public sector data use (Ditchfield *et al.*, 2022), referenced above.

2.6.4 Accountability

In this section, I discuss the concept of accountability regarding algorithmic bias mitigation. As organisations shift more of their human decision making to algorithmic systems, there has been a lack of clarity regarding who is responsible for the harmful or biased decisions produced by these technologies. Leonelli (2016) describes what it means for someone to be accountable, rather than merely responsible, as follows; "responsibility [is] the moral obligation to ensure that a particular task is adequately performed […] accountability denotes the duty to justify a given action to others and be answerable for the results of that action" (Leonelli, 2016).

There has been some debate about whether the algorithmic technologies themselves, or the people who design them, should be held accountable for the outputs of such systems. Floridi and Saunders (2011) claim algorithmic systems have their own moral agency, independent of their designers' moral agency, as these technologies are able to 'learn' and act in a way that is beyond the intentions of their original code. The view of Floridi and Saunders (2011), is these systems should be held accountable for their actions. To put it another way, if algorithmic technologies are assumed to have no moral agency independent of their designers, this may lead to humans being held accountable for algorithmic technologies' decisions. However, despite this, they argue that algorithmic technologies cannot be associated with a sense of moral responsibility or duty (Floridi & Saunders, 2011: Mittelstadt *et al.*, 2016). Algorithmic technologies are designed and implemented by humans. Assigning responsibility to the algorithm can lead to all accountabilities being shifted to the system itself, ignoring the moral responsibilities of the data practitioners and the organisation responsible for its development (Mittelstadt *et al.*, 2016).

Regulatory frameworks can be used as a mechanism to foster accountability for the decisions made by algorithmic technologies. In 2018, the GDPR was implemented across the EU, which included various mechanisms pertaining to algorithmic bias, particularly those relating to the responsibilities of data controllers (GDPR, 2016). Although the UK withdrew from the EU in 2016, the UK is still following the same legal framework. Specifically, article 22 outlines the data subjects' 'right to an explanation' concerning the logics of any fully automated decisions made about them (GDPR, 2016). In Edwards and Veale's (2018) analysis of the 'right to an explanation', they argue citizens are often unable to exercise this right due to the inaccessibility of these mechanisms, meaning these regulations provide citizens with little tangible means of utilising these protections. Edwards and Veale put forward this argument using the example of Data Protection subject access requests (SARs), which they argue have been predominantly used by journalists and company insiders, and not the public at large, due to the amount of time, knowledge, and persistence that following through with these requests requires (p6). Furthermore, if a data subject were to receive an explanation as to how an algorithmic system came to its decision, it is uncertain how this might assist the data subject in receiving a fairer outcome (*ibid*).

Additionally, article 35 of the GPDR stipulates that if the data controller utilises new technologies to process data, and the type of data processing is likely to present a high level of risk to the rights of data subjects, then a data protection impact assessment (DPIA) must be undertaken prior to data processing (Edwards and Veale, 2017). These legal shifts establish mechanisms by which organisations bear some legal accountability for considering the impact of these technologies, in addition to accountability for documenting those considerations (Reisman *et al.*, 2018). However, critics have argued this is insufficient, arguing that while these legal shifts create organisational accountability regarding the production of these documents, such documents provide data subjects with very little practical protection against harm. Moreover, there is no stipulation that those who may be subject to potential impacts of these technologies are consulted in the production of these assessments, prompting concern that the documented impacts will bear little resemblance to the harms faced by the groups of people most likely at risk of harm (Metcalf, Watkins, *et al.*, 2021; Yam and Skorburg, 2021). Further discussion of algorithmic impact assessments can be found in Chapter 6.

2.6.7 Transparency

In this section, I discuss the concept of transparency in relation to algorithmic bias mitigation practices. Within the debate surrounding algorithmic bias, the concepts of accountability and transparency are intrinsically linked to each other. Without knowledge of algorithmic technologies in use, details of how these technologies work, and the working practices surrounding these technologies – that is, transparency - it is not possible for actors to be held accountable for the decisions made by these technologies. In addition, it is argued that transparency around algorithmic systems allows for observation of the mechanics of said system, and so observers will be "better able to judge whether a system is working as intended and what changes are required" (Ananny and Crawford, 2018, p974).

However, a number of barriers exist to creating more transparent algorithmic technology development processes. Organisations may have some incentive to limit the information publicly disseminated about their algorithmic systems – to ensure the information asymmetry between the organisation and the public is maintained. They may be concerned about protecting commercial secrets, user privacy, and ensuring users can't 'game' the system (Veale, 2017). Bates et al's (2023) reflect on the concept of transparency from their work with public sector research partners in the *Living With Data* project, and state that organisations often see information sharing as risky. However, they argue, information which organisations see as being potentially risky may be information the public finds the most relevant and useful regarding systems they may be negatively impacted by.

Additionally, one of the oft cited barriers to algorithmic transparency is that certain algorithmic techniques make it difficult to ascertain how the algorithm has reached its conclusion, and in some cases, there may not be a human-intuitive explanation that can be provided (Strauß and Stefan, 2018: Edwards and Veale, 2017: Zarsky, 2016). As discussed earlier, the move from data modelling to algorithmic modelling acknowledged a trade-off

whereby algorithmic models are less interpretable but gain in predictive power compared to data models (Breiman, 2001). Whilst this may be a less contentious issue when problem solving an engineering issue, it becomes far more challenging when using personal data to make life changing decisions. Furthermore, it is not only the technical understanding of the algorithmic systems themselves which creates a barrier to transparency, but also the structure of the organisation itself. For example, in Bates *et al.*'s (2023) aforementioned paper, they reflect on how practitioners at the DWP and BBC often had contradictory and inconsistent accounts as to the details of their data driven systems. This was due to the interviewees' differing professional roles on the project, with interviewees having different information depending on the remit of their role. These observations highlight the difficulty in organisations providing meaningfully transparent information about their systems.

In addition to these practical considerations, critics have questioned the value of the ideal of transparency in mitigating the potential harms brought about by algorithmic technologies. Ananny and Crawford (2018) have contended that transparency is limited in its ability to defend against corruption; transparency "can reveal corruption and power asymmetries in ways intended to shame those responsible and compel them to action, but this assumes that those being shamed are *vulnerable* to public exposure" (Ananny and Crawford, 2018). This relates to the concept of accountability discussed in the previous section. Without mechanisms to ensure organisations can be held accountable for the development of algorithmic technologies, transparency alone cannot address issues of algorithmic bias. Furthermore, Ananny and Crawford (2018) argue that the ideal of transparency as a mechanism to protect against harm places an unreasonable burden on citizens to seek out information about these systems, to be able to interpret this information, and to recognise its significance. This connects to the issue discussed by Edwards and Veale (2018) earlier, that although article 22 in the GDPR creates a mechanism for the public to make a subject access request (SAR), these requests have been predominantly used by journalists and company insiders, and not the public at large.

Some scholars have further defined the concept of transparency regarding issues of algorithmic technologies. Bates *et al*. (2018) argue it is necessary for organisations to create *socially meaningful transparency practices*, meaning transparency practices which are relevant and useful to multiple and diverse publics. They present the challenges regarding transparency they identified in their research as opportunities for the creation of socially meaningful transparency. For example, in the case of information asymmetry, they suggest that information asymmetry needs to be reduced between the organisations using algorithmic technologies and non-commercial third parties such as researchers, policy makers, journalists, political representatives, service users, and the public. Specifically, they suggest organisations could be legally required to publish what algorithmic technologies they were currently developing. This builds on previous suggestions in the field regarding algorithmic technology registers, such as Amsterdam and Helsinki's algorithm registers (Bates *et al.*, 2023). These registers document algorithms in use in these cities.

Additionally, Bates *et al.* (2018) argue for enhancing 'collaborative governance' of algorithmic systems (Bates *et al,.* 2018; Kaminski, 2020). Collaborative governance of algorithm systems entails organisations working with third parties to decide which parts of an algorithmic system should be made transparent to the public, and the ways in which these transparency efforts ought to be communicated to diverse publics. This proposal aims to address the barrier of information asymmetry by engaging a diverse set of actors in the development of these technologies. Furthermore, Bates *et al.* (2018) argue that enhancing collaborative governance, provides a forum to address the issue of uncertainty within algorithmic technology development. By collaborating with third parties, organisations can learn how to communicate uncertainty regarding a system, and ensure accessibility of their transparency outputs to diverse publics at risk of being negatively impacted by algorithmic technologies. However, it is currently unclear how these types of working practices would be embedded within organisational working practices.

2.6.7 Ethical design

In this section, I discuss the concept of ethics and how it relates to algorithmic bias mitigation. It has been suggested that to encourage a greater sense of moral responsibility amongst data scientists and engineers designing algorithmic systems, there needs to be a greater focus on providing ethical training (Leonelli, 2016). Ethics is understood as the practice of evaluating decisions in regards to moral principles. However, in McNamara *et al.*'s (2018) experimental study, it was found that this may not have an impact on ethical decision-making in practice. The study used ethical vignettes to determine whether bringing attention to the ACM code of ethics impacted software engineers' ethical decision-making (N=168), and concluded this had no impact on the software engineers' decisions compared to the control group (McNamara *et al*., 2018: Green, 2018). Green (2018) argues that data scientists need to go beyond simply adopting ethical and professional codes of practice, to engage in a more reflexive practice where they critically examine the work that they produce. Green (2018) argues this is particularly important considering the political power which data-scientists now wield:

> "data scientists are political actors in that they play an increasingly powerful role […] structuring how institutions conceive of problems and make decisions, data scientists are some of today's most powerful (and obscured) political actors." (Green, 2018, p8)

Despite the power now afforded to data scientists, data scientists often seek to adopt a 'neutral' ethical position, or argue they are unqualified to make political judgements within the context of their work (Green, 2018). This position recalls the idea of the objective gaze from nowhere, and ignores that "[o]bjectivity derives its impetus, and also its shape and meaning, from cultural, including political, contexts" (Porter, 1995, p90).

Ethics is generally understood as being the study of moral principles and how these should guide personal and societal behaviour. Academically, the discipline of philosophy has provided much of the historic groundwork for the different schools of thought in this area. However, in recent years, there has been a surge of organisations adopting what is known as

'organisational ethics' or 'business ethics' approaches to dealing with the moral dilemmas found in organisational working practices (Vogel, 2006; Mckinsey & Comapny, 2022). This has been influenced by the rise of 'corporate social responsibility,' which began in the 1990s, when organisations began focusing more on what types of ethical practice were practical and affordable, in addition to practicing greater 'stakeholder engagement' (Vogel, 2006).

The application of organisational ethics in the context of designing algorithmic technologies has been defined as including both a) the moral considerations surrounding the building of these technologies (e.g. what the technology will do, how transparent the technology is), and b) the moral decisions these technologies will be programmed to make (e.g. if the technology makes a decision which is 'unfair') (Wing, 2018). In Moss and Metcalf's report on Silicon Valley ethics workers (described as 'ethics owners') they describe a shift from the older form of ethics work done in these companies, where ethics workers deflected public pressure and demonstrated legal compliance, to the ethics roles of today which focus more strongly on preventing social harm (Metcalf, Watkins, *et al.*, 2021). Additionally, these roles focus on preventing harm within the parameters set by the business, with issues occurring when ethics workers try and work beyond these limits. For example, Timnit Gebru, in her capacity as an ethics lead at Google, co-authored a paper detailing the risks of large models to exacerbate carbon emissions, have unknowable biases, and spread misinformation (Hao, 2020; Bender *et al.*, 2021). Google managers requested she withdraw the paper from the publication process. After a meeting about how to go forward with this issue, Gebru's employment with Google was subsequently terminated. This type of ethics work differs from academic ethics philosophies due to its strong focus on navigating stakeholder relationships, managing company resources, and the need to navigate market pressures (*ibid*).

One of the challenges of the business ethics framework is the lack of agreement between ethics workers as to what constitutes 'ethics', with different practitioners bringing with them different sets of personal moral codes to their work, making it difficult for ethics workers to create organisational processes which are workable and consistent (Moss and Metcalf, 2020). Moreover, this may lead ethics workers to focus their attention on more quantifiable benchmarks such as bias or fairness measures (*ibid*). Within this paper, I chiefly use the word 'ethics' to describe the processes which organisations use to work towards some form of 'social good' or to otherwise address moral dilemmas within the context of their organisation, such as through ethics boards, procedures and policies.

2.7. Conclusion

In this chapter, I reviewed the literature surrounding algorithmic bias and proposed algorithmic bias mitigation methods. In the first section, I provided a broad overview of 'algorithms' and their history, drawing upon literature from both data science and the social sciences. Using the data science literature, I focused on the disciplinary debate regarding the validity, epistemological assumptions, and best working practices when utilising algorithmic methods. Regarding the literature on algorithmic technologies within the social sciences, I

explored how academics have questioned both the supposed objectivity of these methods, and how they are embedded within social structures.

Following from this, I explored the concept of algorithmic bias itself, and its relationship to the concepts of discrimination and social inequality. In particular, I drew out the tension between the discriminating character of these technologies and attempts to ensure these technologies do not discriminate against individuals. After this discussion, I provided a brief overview of the use of algorithmic technologies within the public sector.

In the final sections, I explored the range of algorithmic bias mitigation methods which have been put forward by practitioners, academics, and institutions focused on algorithmic bias mitigation. These methods included technical de-biasing methods such as statistical parity checks, data sampling methods, and removing variables related to an individual's protected characteristics prior to the use of an algorithmic model. However, in my discussion, I conclude that these methods are not effective in mitigating the risks of algorithmic bias, drawing on critiques from critical data scholars. I subsequently discussed the concepts of fairness, transparency, accountability, and ethics in relation to algorithmic bias mitigation efforts. In this section, I discussed the challenges with each of these methods, and the ways in which these tie in with the broader critiques of categorisation, power, and the trends within algorithmic decision-making plays a part.

While there has been much research around de-biasing techniques using a quantitative lens (Balayn and Gürses, 2021), it has been noted by prominent scholars that further research in this area needs to include qualitative research that can capture "the messy reality of many contemporary on-the-ground situations," (Veale and Binns, 2017, p12). Orr and Davies (2020), Veale *et al.* (2018), and Holstein (2017) have interviewed practitioners to understand how they are situated within the development of algorithmic bias, and their responsibilities and engagements within their working context. However, their analysis is primarily focused on how individual actors are constrained within a collective system (Holstein, McLaren and Aleven, 2017; Veale, Van Kleek and Binns, 2018; Orr and Davis, 2020). Conversely, the way that algorithmic bias mitigation is approached from an organizational or project perspective has received less attention.

In the following section, I take these observations forward, and use them to develop the methodology approach used in my thesis paper.

3.1. Introduction

In Chapter 2, I reviewed the literature around algorithmic bias, and algorithmic bias mitigation methods. In this chapter, I discuss my approach to synthesising this knowledge within my research context and the PhD's overarching research aims. I then discuss how these research aims were turned into research questions, and how I subsequently designed and carried out each research paper.

The structure of this chapter is as follows. First, I expand on how the 3-paper model of my PhD programme worked, and how this influenced the research design process (section 3.2). I subsequently discuss my research problem and how the overarching research aims were broken down into research questions (section 3.3). Then, I discuss my research approach and how this influenced my data collection and analysis (section 3.4). Following from this, I explore the DWP context to provide the reader with information about the research context (section 3.5). Once the research context has been established, I discuss my research design process (section 3.6), my chosen data collection methods and their suitability for this research project (section 3.7), and how the data generated from these projects were analysed (section 3.8). Lastly, I discuss the ethical considerations on this project (section 3.9) and the limitations of my research (section 3.10).

3.2. Discussion of thesis model

In this section, I discuss how I approached the paper-based model used in my thesis. The Data Analytics and Society Centre for Doctoral Training (link here), which funded my PhD, has a strong focus on PhD students collaborating with industry partners. Typically, these industry partners give students a short research brief and provide some support and funding throughout the PhD process. My industry partner was the DWP (Department of Work and Pensions), whose overarching research aim for the project was quite broad – the department sought to use more algorithmic approaches, and wanted to know more about how they might mitigate the risks of algorithmic bias. In response to this brief, my first step was to conduct a literature review and analyse the organisational context, to produce research questions of academic interest whilst practically relevant to my industry partners. In the next section, I further explore how I produced the research questions for my thesis papers.

3.3. Research problem and research questions

As previously noted, the DWP provided me with a short research brief prior to starting my PhD. Additionally, in the early stages of my research, I attended meetings with my DWP contacts to understand what the department's aims were in supporting my PhD. Alongside this, I undertook an internship with the DWP Sheffield data science team. During this time, I spoke to members of the data science team to determine their understanding of algorithmic bias. Often, team members would mention resources about algorithmic bias with which they

had engaged, or tell me about departmental frameworks in development. These resources would discuss, for example, the problems with technical bias checking when a dataset does not include protected characteristics such as gender, which suggested that the team were working with an algorithmic framing of algorithmic bias (as defined in Chapter 2) (Selbst *et al.*, 2018). As I continued to think about my research, these observations were important in trying to ground my understanding of the research context. In other words, these conversations helped me develop a better idea of what "intellectual puzzle" I might investigate within the boundaries of my research brief and the research context. Mason (2002) describes a researcher's "intellectual puzzle," at its most basic level, as "something which the researcher wishes to explain" (p7). Mason (2002) conceptualises this 'puzzle' as the "essence of [the researcher's] enquiry," continuing by stating that research questions are the "formal expression of [the] intellectual puzzle" (p20).

Considering these matters, I decided to design each of my three thesis papers so they would contribute towards my thesis in a systematic manner, whereby the final outcome of the thesis would broadly focus on what DWP data science practitioners could do to mitigate algorithmic bias. Blaikie (2005) describes this type of research as "applied research" – research which typically focuses on change, evaluation, or assessing social impacts. The two types of applied research which seemed relevant to my own were what Blaikie describes as 'change' and 'evaluation.' Blaikie (2005) describes research focusing on change as "[intervening] in a social situation by manipulating some aspects of it, or to assist the participants in doing so, preferably on the basis of established understanding or explanation." Research focused on evaluation meanwhile is described as "[identifying] the social and cultural consequences of planned projects, technological change, or policy actions on social structures, social processes and/or people" (p72).

As previously described, the DWP's research brief indicated a wish to change working practices to mitigate the risks of algorithmic bias. It therefore seemed sensible that the final paper, either through the research process or the research outcomes, should aim to bring about some form of change (Blaikie, 2005). With this end point in mind, it was important for me to find a systematic approach to reaching the final paper's outcomes. Additionally, Blaikie states that to be confident that a change objective is meaningful, it should be based on prior established explanation (Blaikie, 2005). This indicated that prior to the final paper, I would need to better understand the DWP's current data science environment, and to find out which algorithmic bias mitigation methods other organisations have been successfully using. These considerations led to the following overarching research aims:

- *paper one:* to investigate what DWP data scientists are currently doing in areas related to algorithmic bias
- *paper two:* to investigate how other public sector organisations seek to mitigate algorithmic bias
- *paper three:* to investigate how the insights from paper two could be integrated into a DWP context

Additionally, it was important to consider the academic research which had been conducted in this space so far. My review of the relevant literature informed my design of papers one and two. As discussed in Chapter 2, data scientists have developed algorithmic de-biasing techniques, using quantitative techniques that rely on data scientists producing and assessing output metrics, such as error rates and other comparative statistics (Balayn and Gürses, 2021). However, it has been noted by prominent scholars that further research in this area needs to include qualitative research that can capture "the messy reality of many contemporary on-the-ground situations" (Veale and Binns, 2017, p12).

Thus far, there has been little qualitative research regarding data science practitioners' perceptions and experiences of algorithmic bias and algorithmic bias mitigation methods. Orr and Davies (2020), Veale *et al.* (2018), and Holstein (2017) have interviewed practitioners to understand how they are situated within the development of algorithmic bias, and their responsibilities and engagements within their working context. However, their analysis is primarily focused on how individual actors are constrained within a collective system (Holstein, McLaren and Aleven, 2017; Veale, Van Kleek and Binns, 2018; Orr and Davis, 2020). Conversely, the way in which algorithmic bias mitigation is approached from an organizational or project perspective has received less attention.

With this in mind, the aforementioned research aims for each paper were developed into research questions. The final research questions were:

RQ1a: What algorithmic bias mitigation working practices are currently practiced by data science practitioners at the DWP?

RQ1b: What are the limitations of these practices?

These research questions were the focus of *Chapter 4: Investigating the role of current DWP working practices in mitigating algorithmic bias*, and considered how DWP data science practitioners already engage with algorithmic bias mitigation practices, to gain an understanding of what type of "change" may be needed to improve their approach to algorithmic bias mitigation (Blaikie, 2005). In doing so, these questions aimed to find out about DWP working practices in relation to algorithmic bias, to lay the groundwork for the papers in Chapters 5 and 6.

Following from this, the next two research questions were:

RQ2a: What might 'good algorithmic bias mitigation practice' on an algorithmic project look like?

RQ2b: What challenges does good practice on an 'ethical AI' project face in practice?

These research questions were the focus of *Chapter 5: Lessons in mitigating bias from the field: Exploring good practice and moral challenges on the AuroraAI project*. These research questions aimed to find out what other algorithmic projects were successfully using to mitigate the risks of algorithmic bias, so that this could be fed into the final paper in Chapter 6.

Following from this, the next two research questions were:

RQ3a: What aspects of DWP organisational culture might influence the adoption of mitigation approaches?

RQ3b: And, what does this mean for what might work to mitigate algorithmic bias in practice?

These research questions were the focus of *Chapter 6: The influence of DWP organisational culture on the adoption of algorithmic bias mitigation practices and implications for practice*. These research questions aimed to find out what DWP could do to mitigate algorithmic bias.

3.4. Research approach

After defining my research problem, it was necessary to consider my research approach. A research approach typically entails the alignment or development of the researcher's research philosophy. A research philosophy provides the researcher with an approach to handling questions about the purpose of data sources, methods, analysis, and reflexive thinking during a research project (Mason, 2002).

3.4.1 Research philosophy

In this section, I discuss the development of my research philosophy. A researcher's research philosophy is usually understood to include both the ontological and epistemological assumptions about the world presupposed by either the researcher or their research problem[5]. This means the research's foundational assumptions about what the world comprises of (ontology), and how we can know about those things (epistemology). Some academics stress the importance of approaching epistemological and ontological issues separately, although the reasoning for this is often unstated, and the benefits of this approach are thus unclear (Marsh and Furlong, 2002).

However, as Crotty (1998) points out, to what exactly an ontology refers in the case of research philosophy is uncertain. Crotty (2018) states that this is in part due to how "ontological and epistemological [issues] arise together," rather than being separate (p20). To express this another way, while ontological beliefs concern claims about what exists, these claims rest on researchers' prior knowledge, capacity, and ability, to investigate these claims. In other words, ontological claims are strongly reliant on epistemological methods, understandings, and tools. Moreover, it is not clear what is gained by a researcher aligning themselves to an ontological system of beliefs, nor is it clear how this improves the research process. As an alternative to a researcher aligning themselves to a separate research ontology and epistemology, Crotty suggests only using an epistemology, and 'theoretical perspective,'

[5] Personally, I believe researchers can take on different philosophical approaches depending on the nature of their research and the needs of the research problem they are working on. Due to this, I do not assume the research philosophy has to align with a singular and consistent set of philosophical beliefs held by the researcher.

as can be seen in fig. 1 below. The theoretical perspective of each research paper is provided within the literature review of each of my empirical chapters.

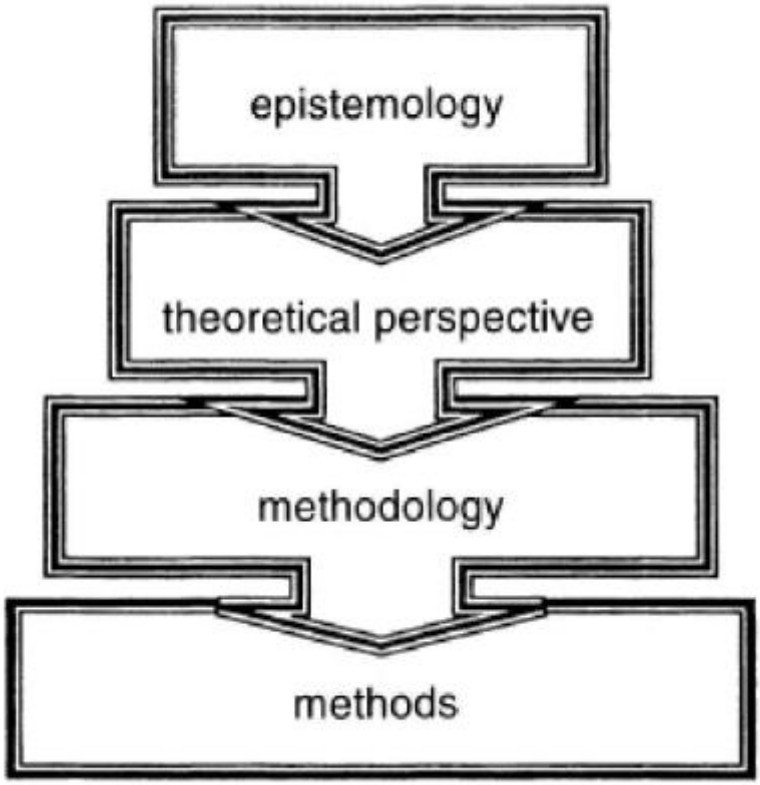


Fig. 1 (Crotty, 1998)

As can be seen in the Fig. 1, epistemological assumptions come first in the researchers' design process, as it is from these that a researcher's broader research area and methodology emerges (Crotty, 1998; Mason, 2002; Blaikie, 2005). Within my own research I found this was a less linear process, where I often iterated on both my epistemology and theoretical perspective in tandem, asking myself questions about the character of my research aims, in addition to the character of the literature from which my research was being developed during the research process. The development of my research philosophy was an iterative process, guided by my findings and the long-term goals of the project. Furthermore, in developing my research philosophy, I did not feel as though I was trying to make metaphysical claims about the nature of reality which are entailed by aligning oneself to an ontological perspective. Rather, I was choosing a way of seeing, understanding, and thinking about a problem which is conducive to studying the problem academically. My focus instead was trying to untangle what my 'intellectual puzzle,' as Mason (2002) calls it, was, and developing my epistemology to match the nature of said 'puzzle.'

As different epistemologies are suitable for different types of research, it was important for me to reflect on the character of what I was trying to understand. During this process, two things seemed key in thinking about my epistemology. Firstly, when reviewing the literature, it was clear that many of the unexplored issues of algorithmic bias were rooted in social interactions, people's experiences, and people's perceptions of the social world around them (O'Neil, 2017; Veale and Binns, 2017; Eubanks, 2018). While some of the more technical research in this area is undertaken from a more positivist standpoint (Hoffmann, 2019; Balayn and Gürses, 2021), there are calls for more research from approaches rooted in interpretivist standpoints (D'Ignazio & Klein, 2020; Veale and Binns, 2017). Thus, to understand the way people experience their social reality, it was important for me to align myself to an epistemological approach which foregrounds the social world. Secondly, it was important for me to adopt a philosophical approach which was complimentary to my theoretical perspective. As my theoretical perspective is largely influenced by critical theory, an interpretivist and social constructivist philosophy were appropriate. I say more about both below.

3.4.2 Interpretivism

As my research questions were focused on the social reality which informs how practitioners engage with data, moral questions, organisational and societal issues, it was sensible to adopt an epistemology which complimented these issues. This led me to study interpretivist standpoints. Modern day interpretivism is:

> concerned with understanding the social world people have produced and which they reproduce through their continuing activities [..] in order to negotiate their way around their world and make sense of it, social actors have to interpret their activities together, and it is these meanings, embedded in language, that constitute their social reality (Blaikie, 2000, p115).

Utilising interpretivist understandings allowed me to consider how practitioners might interpret and construct the social reality in which they work. As Danks and London (2017) have commented, "'algorithmic biases' often cannot be resolved in a purely technological manner, as they involve value-laden questions" (p2). Moreover, they state that these questions are found at every stage of a project, with each organisational context having different values, and policies which reflect those values (Danks & London, 2017, p2). Using an interpretivist lens during my research process led me to considering my participants' work processes from multiple angles – what were the organisation's values? Did participants share them? What influenced these practitioners socially? What thoughts, values, and expectations surrounded these people and the organisation? These types of questions guided my approach. Furthermore, these questions allowed me to hone my conception of algorithmic bias beyond the purely technical, and instead understand it as a living and moving thing whereby each instance of the phenomenon has its own history and hue. This would inform my approach to analysing my research data, as it encouraged me to think about pre-existing biases which may be embedded within the organisation.

3.4.3 Social constructivism

Social constructivism asserts that within our social interactions, we negotiate and construct the meanings of our world and the objects around us. To put it more succinctly, "people act on the basis of the meaning that objects have for them; these meanings are developed through social interaction, and modified through interpretive processes employed in further interaction" (Benton and Craib, 2010; Blumer 1969).

Social constructivism is an epistemological assumption often combined with interpretivism (Creswell, 2012, p8). Creswell states that social constructivist assumptions hold that:

> "individuals seek understanding of the world in which they live and work. Individuals develop subjective meanings of their experiences – meanings directed toward certain objects or things. These meanings are varied and multiple, leading the researcher to look for the complexity of views rather than narrow meanings into a few categories or ideas" (p8).

Indeed, social constructivism is a prominent theme within my broad research area. As described by D'Ignazio & Klein (2020), data collection biases can arise from the way dominant groups conceptualise others, and how the meanings and expectations they have about these people then become a blueprint for how they are treated. However, these processes are not simply prejudices which can easily be uncovered; instead, as Creswell states "[o]ften these subjective meanings are negotiated socially and historically. They are not simply imprinted on individuals but are formed through interaction with others" (p8).

In summary, for this thesis, I developed a research strategy which could be aligned with the DWP's motivations for supporting the PhD. This was achieved through carefully considering the gaps in the academic literature (such as qualitative research on algorithmic bias), and the motivations of the DWP to support research which enabled the changing of their approach to issues of algorithmic bias. In the next section, I discuss the DWP context.

3.5. The DWP context

The DWP (Department of Work and Pensions) is the ministerial government department within the UK responsible for welfare support, pensions, and child maintenance policy. As a ministerial department, it is led by a government minister (who is referred to as the Secretary of State for the Work and Pensions) and supported by a team of junior ministers. Ministerial departments are departments which require direct political oversight. Ministers are drawn from the members of parliament of the currently governing political party, who are elected by their constituents.

Within the UK context, civil service departments stand independently of the government and currently governing political party. Civil servants who work in civil service departments typically stay within the service when new Ministers are appointed when political power is transferred to a new political party following a general election. Officially, civil servants serve

the crown, rather than the government. Due to these factors, while civil servants enact the policies of the governing party, they are considered to be apolitical. The Civil Service Code states that civil servants must demonstrate political impartiality and must not allow their political position to determine any advice they may give or their actions (Civil Servants, Ministers and Parliament, n.d.). In practice, this stipulation requires that civil servants must implement government policy to the best of their ability regardless of their political views, and must not say or write anything which could be quoted in a manner which could be regarded as agreement or disagreement with Ministers' decisions (ibid).

The DWP is the UK's largest civil service department in terms of expenditure, serving around 20 million claimants and customers across pensions and working age benefits (The Audit Office, 2019). Benefits which the DWP oversee include Universal Credit, Jobseekers Allowance, Carers Allowance, Disability Support Allowance, and Personal Independence Payment. The department currently operates over 750 Jobcentres, which provide employment support to those on out of work benefits. Depending on the claimant's circumstances, this support can include weekly check-ins with a designated 'work coach' who assists jobseekers in attaining employment through monitoring their progress on weekly tasks and signposting them to other services (Department of Work and Pensions, 2019).

The employment support provided by DWP has seen a number of changes over the last decade, one of the most noticeable being the development of Universal Credit (UC). The development of UC has meant that what was previously paid as 6 separate benefits (income-related Employment and Support Allowance, income-based Jobseeker's Allowance and Income Support; Child Tax Credit and Working Tax Credit; and Housing Benefit) is combined as the new all-in-one benefit, UC. The roll out of UC began in 2013, and gradually claimants from different regions moved from legacy benefits to UC. This development has also brought with it a move towards a more digitised application system, with UC being an all-digital application service, where claimants must apply for their benefits through the government's online portal.

The UC system has not been without controversy. The shift to online services has caused difficulties, such as claimants without the necessary digital skills struggling to apply for benefits. It is estimated that 6.3 million people (approx. 9.4%) in the UK do not engage with the internet (French, Quinn and Yates, 2018). These users often do not have the skills or support available to them to become digitally active, and they often feel that there is little incentive for them to do so. Non-users are more likely to be from lower socio-economic backgrounds (*ibid*), and thus potentially more likely to be social security claimants. Social security payments are made to citizens in situations such as unemployment, illness, or bereavement and poverty (Mackley *et al.*, 2023). In a UN rapporteur's report in 2018, it was commented the "British welfare state is gradually disappearing behind a webpage and an algorithm, with significant implications for those living in poverty" (Alston, 2018).

Along with the change to an all-digital application service, the policy around Universal Credit includes an increased number of conditions placed upon the recipient of the benefit. To assist claimants in searching for work while claiming UC payments, recipients are required to spend

most of their time performing work search related activities such as looking for work, attending interviews and training sessions, and completing their online work journal. Another feature of the new UC system is that work coaches log the outcomes of their meetings as part of their practice, and digital data about claimants' job search habits when on the UC portal may be recorded to assist DWP in their operations.

Current claimant work search activity requirements build on policies brought in by previous New Labour governments (1997- 2010), which cultivated the concept of 'conditionality' - the principle that state support is dependent on claimants fulfilling certain conditions each week as part of the agreement they enter when they apply to receive state support (Welfare Conditionality, 2019). If claimants do not complete these tasks they can be sanctioned and their benefits may be frozen, thus their benefits are conditional on behaving in the manner stipulated as part of the contractual agreement they enter with the DWP in order to claim social security benefits. Although this move to increased conditionality has been supported by DWP's own evidence-based policy research (Monaghan and Ingold, 2019), which claim these methods are successful at reducing unemployment, increased conditionality has not gone uncontested (Cheetham *et al.*, Moffatt, Addison, Wiseman, 2019; Hardie, 2020).

Dwyer *et al.* (2014) have voiced concern about the increase of conditionality in the UK welfare system. In a qualitative study of benefit claimants over five years, they found that conditionality was more strongly associated with stasis – 'a lack of significant, sustained change in employment status' – than with an improvement in work prospects (Welfare Conditionality Project, 2018). This research contradicts the underlying principle of using conditionality to get claimants back into work. It has been argued that welfare conditionality allows the state to criminalise benefit claimants (Rodger, 2012), and is ideologically aligned with discourses framing those in receipt of benefits as being either the 'deserving' or 'undeserving' poor. As Eubanks highlights in *Automating Inequality*, these discourses serve to individualise the struggles of those in poverty, and minimise their framing as a societal issue (Eubanks, 2018).

Furthermore, the DWP has seen significant changes to policies and services in recent years. After the 2007-2008 financial crisis the Conservative-Liberal Democrat coalition government adopted austerity policies to manage the country's national debt. This has encouraged local councils and government services to seek out data-driven solutions (including algorithmic systems) to reduce costs while managing increasing workloads (Dencik *et al.*, 2018, p120). In 2021, the DWP started trialling an algorithm which detected fraud in Universal Credit claims, with plans to make the algorithm prevent payment of fraudulent claims in advance of payment (Public Law Project, 2022). However, these developments have been the subject of criticism. It has been reported that despite numerous Freedom of Information (FOI) requests, the DWP has yet to provide any information regarding the details of their fraud prevention algorithms (Waterfield, 2023). This has raised concerns about the reliability, legality and fairness of these models (*ibid*). Moreover, due to this lack of transparency, it is not possible for the public to know if these technologies have been found to be discriminatory in practice (*ibid*). Despite these concerns, in 2023 the department committed £70m worth of

investments towards their digital transformation fund to expand their use of algorithmic technologies, some of which is expected to go towards fraud prevention (Waterfield, 2023; DWP, 2023). Relatedly, DWP have committed to generating £1.3 billion worth of savings through their counter-fraud activity in 2023-2024 (DWP, 2023).

As a result of the issues discussed in this section, the investigation of algorithmic bias within DWP is a pressing matter. In the following section, I detail my approach to research design in the DWP context.

3.6. Research design

In this section I discuss how research design decisions were made during my projects, and how these were approached considering that I was working with an industry partner. As discussed in section 3.3, during the early stages of my thesis I planned how each paper would contribute towards the overarching aim of my thesis: to investigate how the DWP might mitigate algorithmic bias.

Following from this, it was necessary to consider the research design of each paper. As discussed earlier, the papers were designed to process from each other in a logical order. In paper one, I explored what DWP data scientists were doing in areas related to algorithmic bias. In paper two, I explored what other organisations were doing to mitigate algorithmic bias. Then, in paper three, I examined how the insights from paper two could be integrated into a DWP context. In the following sections, I discuss the research design process of paper one (section 3.6.1), paper two (section 3.6.2), and paper three (section 3.6.3). Due to the overlap in data collection methods used across the three projects, I discuss my approach to data collection across all three papers in Data Collection (section 3.7).

3.6.1 Paper one

The research questions for paper one were: **RQ1a: what algorithmic bias working practices are currently practiced by DWP data science practitioners?** and **RQ1b: What are the limitations of these practices?**

To investigate these questions, it was important to consider both how I would discover the focus of current bias mitigation, and how practitioners engaged in those areas. These two questions formed two distinct stages in my data collection process. First, I sourced suitable bias mitigation documents from outside the DWP and analysed these documents to understand their content. I then created an interview guide based on the themes in these documents. Secondly, I interviewed practitioners about these areas, in relation to a DWP project they had recently been working on. Sample questions from the interview guide used in this project can be found in section 3.6.1: Interview Guide, below.

When designing this study, I considered the most suitable unit of analysis to investigate my research questions. In this case, I chose to focus my study on a single data science project which was active in the DWP data science team. While I considered interviewing practitioners about their working practices more generally, i.e. without reference to a specific project,

there were several reasons why I did not adopt this approach. To begin with, after reviewing the literature there were already studies which had taken a similar approach – such as interviews with data scientists undertaken by Veale (2018) and Orr and Davies (2020). While these studies investigated how data scientists experience algorithmic bias at an individual level, the findings still left further questions about the social dynamics surrounding algorithmic bias in an organisational environment. Considering that data science projects are often worked on as part of a team embedded within a larger organisation, it seemed prudent to take an approach which allowed for analysis of the dynamics occurring at project level. I therefore chose to take a project focused approach, to understand how decisions are made on DWP data science team projects that involve claimant data, to change DWP services.

Originally, paper one was designed to have two rounds of interviews. The first round would be semi-structured interviews and would involve me speaking to people who had managed the project, those who worked as policy analysts at DWP, and Jobcentre work coaches, to obtain an understanding of goal setting and communication practices on the project. Afterwards, there was to have been a second round of interviews with data scientists who had coded on the project, which would have used elicitation interview techniques. These interviews would have focused on the participant's code as an elicitation device to understand their thought processes during the project. However, due to the Covid-19 outbreak, this planned approach was not possible. After two interviews for the project, Covid-19 was declared a pandemic. After discussion with my DWP contacts, we agreed I would conduct the next interview online. Instead of the elicitation interviews, I interviewed the data scientists about their work on the project in relation to commonly identified areas in mitigation practice.

In addition to changing my research methods, the outbreak caused disruption to my working practices, and to the operation and working practices of those at DWP. These changes meant some participants were no longer able to be interviewed, as they were either in a face-to-face role and unable to attend an online interview, or they were re-deployed to the DWP's frontline services to manage the influx of unemployment claims.

PROJECT SELECTION

At the time of conducting the paper one fieldwork, the data science team were in the process of developing two projects suitable for study. However, one of these projects was still in its infancy, and after careful discussion with DWP contacts it was decided the project would not produce enough data for appropriate analysis. The second potential project was the Digital Trialling Framework (DTF) and the Digital Plus Trial (DPT). The DTF and DPT projects looked at comparing the performance of claimants, work coaches, and Jobcentres when an online version of UC's job seeking service was used in place of the traditional face-to-face service the DWP were using at the time this research was conducted. This project also had the benefit of involving multiple sections of the DWP, including data science, policy, and work coaches. While these two projects, DTF and DPT, were technically separate, their management and implementation was intertwined. The projects were jointly run by the DWP Sheffield data

science team and the Labour Market Policy Group (a specialist policy analyst team focusing on labour market solutions at the DWP). The sampling criteria for my research on this project was people who had worked on the DTF or DPT projects. More about sampling on this project can be found in the section below, and further description of the DTF and DPT projects can be found in Chapter 4 (paper one, Methodology).

SAMPLING

The sampling criteria for paper one was that the individual had to have worked on either the DTF or DTP project, and to have worked at the DWP for at least six months. Mason (2002) stresses the importance of ensuring that the research sample includes a 'meaningful range' of participants. For this project, I understood a meaningful range to mean speaking to as many of those who had worked on the project as possible, from each section of the DWP who had worked on the project.

I interviewed six participants for this project, with 2 being interviewed twice due to their role managing the data science team. Of these six, four came from the data science team and the other two came from the Labour Market Policy Group, and acted as the projects' clients. While a small sample size, this included most of the data science team members who had worked on the project. Participants were identified through established contacts at the DWP.

INTERVIEW QUESTIONS

For this study, the research questions were derived from careful document analysis of algorithmic bias frameworks. To create an interview guide, I searched through these frameworks to find common themes. For more detail about this process, see section 3.7.3: Document Analysis.

As my interviewees were from different backgrounds, they had different areas of specialism on the project, so not all interviewees were asked the same questions. A selection of the questions asked can be found below:

- What were the project's aims when it started?
- What stages are there that a project goes through at DWP?
- How do you feel the project's changed since you started on the project?
- Where has the data from the project been collected from?
- What sort of documentation came with the project data?
- How has sampling been handled as part of the project?
- Were there any concerns about bias on the project? … Could you give an example?
- How have developments in the project been communicated to outside stakeholders?
- What do you feel has been the most prominent consideration when it came to designing this project?
- What do you think was the greatest challenge when ensuring the project was fair?
- What sort of regulatory considerations have arisen during the project?

Participants were also asked specific questions depending on their role on the project.

3.6.2 Paper two

The research questions for paper two were **RQ2a: What might 'good practice' on an algorithmic project look like?** and **RQ2b: What challenges does good practice on an 'ethical AI' project such as AuroraAI face in practice?**

After my experience with Paper one, I wanted to continue using a project as my unit of analysis, to provide me with data on the social mechanics at play on an algorithmic project. However, to continue with this approach on Paper two, I needed to find a project in which practitioners had attempted to implement something which looked like 'good algorithmic bias mitigation practice.' Thus, my first step was to identify a suitable public sector project which looked suitable. The project chosen was the AuroraAI project, run by the Finnish Ministry of Finance. I describe the process used to select a suitable project below.

PROJECT SELECTION

To select a suitable project, I conducted desk research to find out what algorithmic projects were being developed in the public sector. The objective was to find a public sector project taking strong steps to mitigate the risks of algorithmic bias. The criteria for this were as follows. First, the project must have publicly posted their ethics guidelines, which had to include consideration of algorithmic bias. Furthermore, as the findings from this paper would form the basis for paper three (in which I would seek to integrate 'insights from the project selected for Paper two into a DWP context), it was important for the project to concentrate on a work or welfare problem. Additionally, the project had to be at a stage where some technical development had started to take place. Other criteria included the need for the project to process data relating to a person, due to algorithmic bias being a phenomenon which develops during the processing of personal data. In addition to these considerations, there were also issues of accessibility, since the only researcher on the project, I would only be able to interview practitioners in English. To look for a project which fit these criteria, I used a combination of internet search methods in conjunction with the Oxford AI readiness index – an index which assesses how ready governments' are to use AI in their public services – to pick countries which were more advanced in this area (Oxford Insights, 2019). Additionally, I searched public algorithmic registers such as the Helsinki algorithm register. These combined criteria proved very restrictive, and left the most suitable project as the AuroraAI project by the Finnish Ministry of Finance, a description of which can be found in Chapter 5.

RESEARCH DESIGN

To answer my research questions, I utilised two research methods; document analysis and semi-structured interviews. I describe the use of these methods in further detail in sections 3.7.3 (document analysis) and 3.7.1 (interviews).

After interviewing practitioners on the AuroraAI project and reviewing the data generated, I had not satisfactorily answered my research questions. The AuroraAI project had experienced unforeseen difficulties, and there had not been time for the team to implement much of the ethical guidance they had drawn up. For this reason, at the time of the interviews my participants' experience with the practical implementation of the project's algorithmic bias mitigation methods lacked the detail necessary to answer my research questions. Furthermore, the size of the AuroraAI project, combined with disparate organisation and communication methods, had made it difficult to know what had been trialled, and who this had been reported back to. This posed difficulties, and there was a need for me to re-evaluate my research project at this stage. It had already been quite difficult to find a project which appeared to fit my criteria, and with the remaining time it would be difficult to find knowledgeable participants working on similarly progressive projects.

Ultimately, I decided to widen my sample to include those working in organisations fighting against algorithmic bias, and those in algorithmic discrimination focused roles more generally. I hoped they would be aware of 'good practice' projects and how these projects were looking to mitigate the risks of algorithmic bias. To ensure continuity through both sets of interviews, the algorithmic justice experts were asked questions relating to both the AuroraAI algorithmic bias mitigation plans, as well as about projects or algorithmic bias mitigation methods which they were aware of and knew to be successful.

SAMPLING

For paper two, I interviewed six participants who were working on the AuroraAI project. These participants had roles ranging from project co-ordination, data engineering, data work, through to ethics expert. Participants were approached in various ways, including emailing key figures mentioned in the AuroraAI documentation, posting an invitation on the AuroraAI slack channel, and snowball sampling. In snowball sampling, the researcher contacts a small group of individuals who are relevant to the research project, and then through these established contacts, further contacts who fit the sample criteria can be reached (Bryman, 2004).

As previously described, my priorities changed during the project on discovering that the AuroraAI interviews would not give me enough data to answer my research questions. In the second round of interviews, I interviewed seven participants who worked in organisations fighting against algorithmic bias, and those in algorithmic discrimination focused roles more generally. Potential participants were contacted based on their engagement with algorithmic justice or data ethics focused work. Organisations which were contacted included the Algorithmic Justice League (USA), Data Justice Lab (UK), Algorithm Watch (EU), and the Ada Lovelace Institute (UK).

INTERVIEW GUIDE

An interview guide was created prior to the interviews, informed by textual analysis of the project's public facing documents (the process of which can be found in section 3.7.3: Document Analysis). A selection of the questions asked to participants working on the AuroraAI project can be found below:

- Could you tell me a little bit about your role on the project?
- Aurora uses a 'human centric' design model – can you tell me a bit more about that?
- The Aurora documentation mentions the importance of citizen input and co-design principles; how has this been incorporated into the design process so far?
- Could you describe the process for design decisions on the project so far?
- What steps have you taken towards implementing the design of the project? … have there been any unforeseen challenges which weren't accounted for in the design stage? … could you give an example of this?
- How have the design team and implementation teams communicated during the preliminary trials? … could you give an example of a challenge that has come up translating the design plan into the implementation stage?
- What sort of ethical guidelines have been drawn up so far? … how have these been incorporated into the design process for Aurora?
- Can you describe the design process put in place to ensure Aurora doesn't discriminate against marginalised groups?
- Have there been any groups which have been of particular concern during the design stage? … could you tell me a little bit about how they were identified?
- How have concerns about algorithmic bias been handled in the preliminary trials?
- What lessons do you think have been learnt so far from AuroraAI's work on fairer AI, that could be applicable to public sector organisations in other countries?

For the round of interviews with algorithmic justice focused participants, participants were asked questions relating to the AuroraAI's projects ethics plans to date, and were also asked whether they knew of any algorithmic public sector projects which had been successful in mitigating algorithmic bias. A selection of the questions asked to these participants can be found below:

- Could you tell me a little bit about what you do/did at [insert organisation]?
- What's your understanding of what algorithmic bias is?
- What is your opinion of the suggestion that transparency can make AI less biased?
- How can data and AI transparency be improved?
- Are you aware of any examples/projects which 'do transparency well'? If yes, please tell me about it/them.
- The AuroraAI R&D process has identified co-design as a route to its ethical implementation and mitigating the impact of algorithmic bias. What is your opinion of the suggestion that co-design can make AI less biased?
- How can data and AI co-design be improved?

- Are you aware of any examples/projects which 'do co-design well'? If yes, please tell me about it/them.
- The AuroraAI R&D process has identified value-led design as a route to its ethical implementation and mitigating the impact of algorithmic bias.
- What is your opinion of the suggestion that value-led design can make AI less biased?
- How can data and AI value-led design be improved?
- Are you aware of any examples/projects which 'do value-led design well'? If yes, please tell me about it/them.
- Do you know of any examples or projects which are doing a good job of attempting to mitigate algorithmic bias? If yes, please tell me about it/them.
- What aspects of [the example] do you think could be applied elsewhere?
- What might the design process for an algorithmic system which effectively mitigates algorithmic bias look like?
- What advice would you give to public sector organisations, particularly in work and welfare, in creating algorithmic systems which effectively mitigate the risk of algorithmic bias?

3.6.3 Paper three

This study addressed the following research questions: **RQ3a: What aspects of DWP organisational culture might influence the adoption of mitigation approaches?** and **RQ3b: And, what does this mean for what might work to mitigate algorithmic bias in practice?**

As the final paper in my thesis, this was where all three research projects would come together. In paper one, I scoped out the DWP context in relation to algorithmic bias mitigation frameworks. In paper two, I investigated 'good practice' algorithmic projects, and the pitfalls they might face while attempting to mitigate algorithmic bias. In paper three, I explored how the knowledge learned in paper two could be interpreted, used, or be beneficial in a DWP context.

Earlier, I discussed how Blaikie identified *change* as a potential research objective, either through the research itself or as a result of the project's research outcomes (Blaikie, 2005). Indeed, this is also an objective of 'action research' (*ibid*). Action research paradigms aim to bridge the gap between theory, research, and practice (Holter and Schwartz-Barcott, 1993). While my paper three project did not work within an action research paradigm, this paradigm provided inspiration which respect to the type of methods I might use in this paper. I therefore chose to use workshops as a data collection method in my third paper. Further discussion about the theoretical considerations of this method can be found in section 3.7.6: Workshops.

The research design for this study came in two parts. During the first part, I held seven educational workshops at the DWP. The first six workshops were designed as a multi-part series, the content of which can be found in section 3.7.6 (Workshop methodologies). The first 4 workshops were designed to introduce participants to different concepts in algorithmic bias mitigation, with the final two workshops providing attendees a chance to think about

how these approaches might be implemented within a DWP context, and to start to develop their own algorithmic bias mitigation framework suitable for a DWP context. After the original series of six workshops, a participant approached me about running an extra session for another group within the DWP, due to growing organisational interest in this area. After discussing this together, we decided what would be the best material to include in this two-hour session.

WORKSHOP SAMPLING

Participants self-selected onto the workshop series, and information about the workshop series was spread via word of mouth. My DWP contacts approached people in the organisation who they thought would be interested, either through individual or group emails. These emails contained a small summary of the contents of the workshop. Furthermore, established contacts at DWP offered to assist in scheduling the workshops, to ensure the workshops fitted into potential participants' working schedules. Participants were encouraged to come from a wide range of backgrounds and knowledge domains within the DWP, such as data science practitioners, work coaches, diversity specialists, and research co-ordinators. The only criterion for attending was an interest in algorithmic technologies and discrimination. Approximately 30-35 people attended the workshops.

INTERVIEW SAMPLING

After the workshops, participants were invited to take part in follow up interviews. Anyone who had come to one of the seven educational workshops was eligible for the follow-up interviews. Seven participants attended a follow-up interview, which took place between 1-3 months after the workshops, depending on the participant's schedule.

INTERVIEW GUIDE

Participants were asked questions about their experiences during the workshops, such as which sections of the workshops they found useful, and which they found less useful. Participants were asked questions which asked them to recall specific examples of when they had thought back to the workshop material, to ground answers in the participants' situational experience and the specifics of their own lived experience (Mason, 2002). The questions asked were as follows:

- What's your role in the department?
- What did you know about algorithmic bias before attending the workshops?
- What was your main takeaway from these workshops?
- Have your thoughts on attempting to mitigate algorithmic bias changed since watching the workshop series? … in what way?
- Which parts of workshop content did you find the most useful in the context of your work with DWP…. Why?
- Which parts of workshop content did you find the least useful in the context of your work with DWP…. Why?

- Since attending the workshops, has there been any time where the content of the workshop series felt particularly salient to your everyday working practice? …Can you give an example of this?
- Did you feel anything changed in how this situation was approached? … how so?
- Do you think there are limitations mitigating algorithmic bias in a DWP context?
- [Workshop 2] The workshop involved exercises where you considered the values of different stakeholders, and how their concerns and needs might be in tension… what are your reflections on these sorts of exercises?
- [Workshop 3] The workshop involved doing an algorithmic impact assessment with an ethical vignette… what are your reflections on this exercise?
- [Workshop 4] The workshop involved an exercise with a dataset, focused on better understanding what you might know or not when looking at data… what are your reflections on these sorts of exercises?

3.7. Data Collection

I will now detail my approach to data collection, and the methodological considerations which were important during this process. I will cover each method in turn and discuss how these were used in the relevant papers. In section 3.7.1, I discuss the different types of interviews I used (semi-structured interviews, elicitation interviews, expert interviews). In section 3.7.2 I discuss document analysis. In section 3.7.3, I discuss workshops.

3.7.1 Semi-structured Interviews

The semi-structured interview format was used in all three research papers. The structure of semi-structured interviews falls between unstructured and structured interviews. In structured interviews, a researcher has a set list of questions they ask the participant, and only these questions are asked. Therefore, the interview is fully structured, and does not allow for exploration outside of those questions. Burgess (1984) comments;

> "[i]t is assumed that the interviewer can manipulate the situation and has control over a set list of questions that have been formulated before the interview and which are to be answered rather than considered, rephrased, re-ordered, discussed and analysed. In short, the interviewer is assumed to have power over the respondent who is given a subordinate role in this context." (p. 83).

Burgess frames the reasoning for this as being one of control, where the interviewer is attempting to extract systematised data during the interview process. In doing so, this style of interviewing shares similarity with the data collection process of surveys, which aims to collect data which fits into well-defined categories. However, it has been suggested that interviews are different, due to the relationship between the interviewer and participant. In this relationship, knowledge is re-constructed through the engagement between the interviewer and participant (Mason, 2002; Kvale, 1996).

In contrast to the structured style of interviewing, an unstructured interview is one where the researcher has no preplanned questions, and the interview is more exploratory in nature.

Burgess (1984) describes this type of interview as being more conversational than that of structured interviews. In unstructured interviews, the interviewer can spend more time focusing on making the interview process more pleasant for the participant. This type of interview can thus be seen as focusing on making the interview "pleasing to the persons interviewed. It should seem to him or her an agreeable form of social intercourse" (Webb and Webb, 1932, p. 139).

In addition to being more agreeable experience to participants, unstructured interviews can produce data which allow for unexpected themes, greater depth, nuance, and complexity, due to the way knowledge is constructed in these interviews when compared to survey methods (Mason, 2002). Unlike structured interviews, unstructured interviews leverage the relationship between the interviewer and participant, allowing for the participants' knowledge to be reconstructed during the interview process (Mason, 2002). The interviews were semi-structured, with questions and topics being prepared in advance and adjusted to each participant, to fully leverage what the qualitative interview offers (Mason, 2002).

In a semi-structured interview, the researcher has a set list of questions, but will however ask follow-up questions which are tailored to the participants' answers and experiences. This type of interview has the benefits of structured interviews, in that the conversation between the participant and researcher is directed towards answering the researcher's research question. However, it retains some of the benefits of unstructured interviews, as it allows for the researcher and participant to explore the topic in a way which is individual to the participant's experiences in this area. Burgess describes this style of interviewing as "conversations with a purpose" (Mason, 2002; Burgess, 1984, p102).

As my research questions focused on the social reality informing how practitioners engage with data, moral questions, organisational and societal issues, it was sensible to pick a research approach and methods which enabled me to address these issues. Interviews are considered a suitable way of collecting these types of insights, as they are complementary to interpretivist approaches, and can illustrate how something is done or how it is experienced (Mason, 2002; Brinkmann, 2013).

Due to the outbreak of Covid-19, I needed to quickly pivot from face-to-face interviews to ones using platforms such as Zoom or Google Meet. Interviews were held using video call platforms, with some interviews having a follow up discussion via email. Interviews lasted between 40-120mins.

3.7.2 Expert interviews

Mason (2002) stresses the importance of a qualitative research sample following from the researcher's theoretical understandings (p122). The people the researcher interviews must have something to bring to bear on the researcher's intellectual puzzle. In the case of algorithmic bias, it is particularly important to interview the experts whose worldviews contribute towards the models and algorithms created. Cathy O'Neil, an expert data scientist herself, suggests that models are simply "opinions embedded within mathematics" (O'Neil,

2017). In this statement, O'Neil not only questions the supposed 'objectivity' of mathematical models, but also draws into the scope of enquiry the individuals from whom these opinions come.

Following on from my social constructivist ontology, those involved in the social construction of algorithmic bias it would be important to interview. Expert interviews are not conducted to better understand the experts' field of expertise – as this can be found in textbooks or other less time intensive sources – but to understand the social practices and institutions which experts affect and move within (Bogner, Littig and Menz, 2018).

Across the three research projects, interviews were often with experts, some of whom had PhDs and some of whom were in management roles. Due to the position which these participants had within their organisations, these interviews can be understood as expert interviews. Expert interviews are useful for understanding how the framing of particular problems might be influenced, and the sense making which goes into the social construction of issues relating to expertise (Bogner, Littig and Menz, 2018). Within the three studies, these interviews allowed me to understand the processes and contextual knowledge these individuals were engaged with, to understand a) the bureaucratic and working practices which surround the spaces where algorithmic bias might develop, and b) how the context, power, and structure within the organisation may stifle – or enhance – efforts at mitigating such bias (Van Audenhove and Donders, 2019).

3.7.3 Document analysis

Document analysis was used in all three of my papers. Project documents not only contain important factual information but also 'social facts' – that is, descriptions of how they are organised, shared, and the differences which can be found between them (Bowen, 2009). This type of evidence is particularly valuable to project-focused studies, due to its ability to provide rich details about the project under investigation, and to uncover new meanings throughout the document analysis process (Bowen, 2009). Not only is this method useful for allowing a researcher to identify important aspects of a project, but also when considering these documents from an interpretivist standpoint, they provide important clues as to how the social life of a project is constructed (Corbin & Strauss, 2008; Strauss & Corbin, 1998; Bowen, 2009).

During papers one and two, this method was used to develop interview questions, and to provide contextual insight into the issues at hand. In paper three, the documents produced during the workshops were then later analysed as secondary sources (Ørngreen & Levinsen, 2017).

3.7.4 Paper one

Document analysis was undertaken in paper one to inform the interview schedule. To find suitable documents, I searched for currently available algorithmic bias mitigation guidance documents in academic, grey, and commercial literature which focused on algorithmic bias

mitigation. This search was carried out through a combination of Google Scholar, Web of Science, and internet search sites such as DuckDuckGo and Google. During this search, my criteria for inclusion were: a) the document looked as though it could be translated into somewhat practical advice by an organisation, b) it appeared to fit within the DWP data science context, and c) it was aimed at an organisational level which would be relevant to my participants. Examples of topics which were considered suitable included: suggestions as to what data science teams should be doing to avoid algorithmic bias, techniques which may or may not be favourable, and guidance for setting out a plan for future work in an organisational context. Three guidance documents were selected as being relevant to the DWP context. These included;

- PwC's responsible AI toolkit (Rao, Palaci, and Chow, 2019)
- The Centre of Data and Ethics Landscape Summary on Algorithmic Bias (Rovatsos, Mittelstadt, and Koene, 2019)
- Cramer *et al.*'s framework for algorithmic bias (Cramer *et al.*, 2018)

The PwC Responsible AI toolkit provides high-level management advice around developments in AI, focusing on fairness, governance, security, and organisational workflows. The Centre of Data and Ethics Landscape Summary draws together academic literature and public sector concerns to provide an overview of algorithmic bias and some potential solutions. These potential solutions include technical bias checking, procedural fairness methods, and project management. Finally, the Cramer *et al.* framework to algorithmic bias examines mitigating bias within the context of an organisation in the music industry. This focuses on how to assess algorithmic bias in practice, and how to communicate about it across different teams.

Once I had selected these documents, I analysed them for themes to inform my interview guide. During this process, I noticed there were two prominent types of advice within the documentation, technical and managerial advice. The themes developed within this stage of the project were also used as codes during the data analysis, to guide analysis of the guidance documents.

3.7.5 Paper two

In paper two, documents were analysed to understand the AuroraAI project. To find suitable documentation, I looked through the AuroraAI project's public facing documents. Documents available on the AuroraAI website include:

- A 2-hour video of the AuroraAI International Conference (in English)
- A 7-hour podcast on themes about a human centred society and AI (in Finnish)
- Mission statement documents available on the Finnish Ministry of Finance website (available in English)
- The development and implementation plan (2019–2023) based on the preliminary study on the Aurora national artificial intelligence programme (43 pages, in English)

These documents were analysed to produce the interview guide for the empirical research for paper two, and also provided supplementary data to analyse alongside the interview data. The first two podcast episodes mentioned above were translated into English for analysis. However, due to unforeseen circumstances, the translator was unable to translate any of the following episodes or documents, and consequently, only two hours of the podcast series were analysed.

3.7.6 Workshops

Research paper three used workshops as a data collection method. Workshops are a kind of action orientated research method (Freytag and Young, 2017). Like focus groups, workshops include interaction between participants, and allow opinions to be revealed which might not otherwise have surfaced in a traditional one to one interview (Morgan, 1998). This method is thus well suited to sociological topics, as it provides a window into how social construction plays out in interpersonal relationships.

In addition, the workshops provided a space where practitioners of different types could create a shared understanding of the issues they were discussing (Morgan, 1998). This was a very important consideration, as the findings of the second paper suggested that some of the difficulty in mitigating algorithmic bias came from a lack of shared understanding between practitioners. Using workshops as a research method created a space where participants could jointly identify and articulate language around 'fuzzy' issues, and also develop a shared sense of understanding and communication (Ørngreen & Levinsen, 2017). These types of spaces also allow for tacit knowledge to become more evident, as participants are required to communicate with each other about their own workings and assumptions (Freytag and Young, 2017). This was also the basis for my decision to have the workshops open to a wide selection of DWP workers. The workshops were open to anyone across data science, policy, or operations sections, as long as they were interested in algorithmic developments.

The educational material for workshops 2-4 was developed out of paper two's findings. During the analysis of paper two, I identified algorithmic bias mitigation methods which would be suitable for inclusion in the workshops. These included Value Sensitive Design (VSD), algorithmic impact assessments (AIAs), and critical thinking skills, which are further discussed in Chapter 5. Once these were decided upon, I planned a workshop around each of these methods, by reviewing the literature in these areas and looking for activities which had already been designed in relation to those methods.

For the VSD workshop, I included a direct and indirect stakeholder analysis exercise, as well as a value source analysis exercise, to provide an introduction to the VSD framework (Friedman, Hendry and Borning, 2017). For the algorithmic impact assessment workshop, participants were asked to complete the Canadian Government's Algorithmic Impact Assessment Tool online using information from a vignette (Treasury Board of Canada Secretariat, 2021). In the data empathy workshop, participants were given a public sector dataset to explore, and were asked to explore their own assumptions about what the dataset was about, and who it belonged to. Session 5 involved participants rapid prototyping an

ethical algorithm, a process which involved identifying the algorithm's scope, what data would be used, and what the business case was for this algorithm. The activities in this session repeated some of the exercises in sessions 2-4. In session 6, participants were asked to read through the CEDI's data ethics framework and discuss their thoughts on it. The slides and activity Jamboards for the workshop can be found in Appendix A and B. In Workshop 7, I gave a compressed talk covering the content of the six workshops in addition to the algorithm prototyping exercises in Workshop 5.

No.	Workshop session content	**Length**	**#participants**
1	**Workshop 1: Introductory session (WS1).** This session involved a talk on projects 1 and 2, as well as an introduction to project 3. 30mins Q&A and discussion.	1 hour	15 (Approx.)
2	**Workshop 2: Designing with values in mind (WS2).** This session involved activities based on value sensitive design methods, focused around stakeholder analysis and engagement.	1 hour	8 (Approx.)
3	**Workshop 3: Impact assessments for data-driven technologies (WS3).** This session allowed participants to explore new impact assessment standards being developed to mitigate the impact of bias in data-driven technologies.	1 hour	6 (Approx.)
4	**Workshop 4: People behind the datasets (WS4).** This session focused on the people behind the numbers, using creative story-telling exercises and ethical vignettes to explore issues of bias.	1 hour	7 (Approx.)
5	**Workshop 5: Algorithm prototyping session (WS5).** This session focused on prototyping an algorithm which effectively mitigates risks of bias.	2 hours	7
6	**Workshop 6: Framework prototyping session (WS6).** This session involved participants creating their own framework for mitigating bias when using algorithmic technologies at the DWP.	2 hours	2
7	**Workshop 7: Compressed workshop series session (WS7).**	2 hours	18

(Fig., B, see Appendix C or a fuller draft of each workshop)

Workshops were held online, using Microsoft Teams, as this was DWP's preferred choice of platform. In addition to Microsoft Teams, participants were provided with Google Jamboards in each session, which allowed them to engage with the session's activities using digital post-it notes. The data generated from these workshops was of two types: One, the workshops

were video recorded. Two, workshops often involved some type of written or visual activity produced as part of the activity.

3.7.7 Elicitation interviews

Prior to the outbreak of Covid-19, I planned for a second round of interviews on paper one, which aimed at understanding practitioners' thought processes when using quantitative methods. These interviews would have been face-to-face, and would have involved interviewing practitioners about the code they had developed as part of the project. This method of interviewing was piloted in 2018, and appeared to be useful at gaining insight into the thought processes quantitative practitioners have whilst working. Sadly, due to the Covid-19 pandemic, these interviews were not possible. However, here I will explain the original plan.

Theoretically, the premise for these interviews was based on artefact elicitation interviews. Artefact elicitation interviews, much like the photo-elicitation method on which they are based, have been used in engineering educational research to provoke deep and meaningful responses which situate the participant within their decision-making processes (Douglas, 2015, p1). Like photo-elicitation, this method focuses the interview on a visual artefact, often one created by the participant themselves. This project examined how an artefact interview might be beneficial to understanding the work that data scientists engage in, and uncovering often tacit and situational knowledges relating to the process. Harper comments that "[p]hoto elicitation may overcome the difficulties posed by in-depth interviewing because it is anchored in an image that is understood, at least in part, by both parties" (Harper, 2002, p20). This method was to be used to help overcome the boundaries found when discussing abstract technical concepts, as well as mimicking the way in which programmers discuss their work when speaking to colleagues.

3.8. Data analysis: Thematic analysis

After conducting the interviews and workshops, audio files of the interviews were used to transcribe the data into text format. The transcription process was different on each project. In the first project, I transcribed all interviews without assistance. In the second project, a professional transcriber transcribed the interviews. In the third project, transcripts were created using a combination of the Teams auto-generated transcripts function, combined with personally reviewing the transcripts and correcting them. The reason my transcription process changed between projects was primarily due to the disruption of Covid-19, which left me with less time to transcribe than was previously anticipated.

Once transcribed, the interviews were analysed using thematic analysis of the data (Braun *et al.*, 2014). Braun and Clarke have recommended the following process of thematic analysis; 1) data familiarisation and writing familiarisation notes; 2) systematic data coding; 3) generating initial themes from coded and collated data; 4) developing and reviewing themes; 5) refining, defining and naming themes; and 6) writing the report (Braun and Clarke, 2021).

Step 1 was performed by printing the transcripts and making notes on them while reading. During this phase, I listened to the audio recordings to get a better sense of the conversation and meaning expressed through intonation and verbal gestures.

The systematic coding of the data, Step 2, was broadly the same for each project. Each transcript was coded utilizing NVivo, using a combination of both deductive (codes which were decided prior to the coding process) and inductive codes (codes which I developed during the analysis process). Deductive codes were drawn from questions in each project's interview guides. For example, for project 1, deductive codes included 'technical bias checking', 'governance', and 'project communication', as the interview guides contained questions regarding these issues. The rationale for this was to make it easier to analyse the data in relation to which questions the participants were asked, and to allow me to 'see' more quickly the link between the theory I had relied upon to conduct the research and my data. After this initial coding phase, the rest of the coding was done through the generation of inductive codes.

When generating inductive codes, Mason recommends that this is a time to "direct your attention back towards ontological and epistemological matters" to ensure the coding system is consistent with the philosophical assumptions made during the research design process (p150). As my research was based in an interpretivist philosophy, part of my data analysis process involved taking interpretive readings of my transcripts. This would "involve […] constructing or documenting a version of what you think the data mean or represent, or what [I thought I] could infer from them" (Mason, 2002, p149). The way I grouped and generated the names for my inductive codes varied. Some were things said by participants, such as regularly used phrases. Others were values, concerns, hopes, or emotional aspects. Some were regular talking points, for example, dystopias and utopias.

Traditionally, interpretivist analysis follows the principles of grounded theory (Corbin & Strauss, 2015). Using this framework, researchers are meant to avoid having pre-conceived notions about the data they are analysing. It is worth noting that this ideal is near impossible in practice, as researchers needed to formulate research questions prior to conducting research. While my use of deductive codes was not inline with some interpretivist analysis philosophies, this decision was made because during the course of my PhD, I have studied the issue of algorithmic bias in depth, meaning it would seem disingenuous to suggest my coding process could be purely inductive. Furthermore, the design process for each paper involved gaining familiarity with the organisations and their working practices prior to writing my interview guides. For this reason, I undoubtedly had my own preconceptions and experiences of the data collected prior to analysis. Having those deductive codes available to me while analysing the data made my own theoretical assumptions more explicit, providing me with further opportunity for reflexivity while analysing my data. Coffery and Atkinson state that "[t]heories are not added only as a final gloss or justification; they are not thrown over the work as a final garnish. They are drawn on repeatedly as ideas are formulated, tried out, modified, rejected, or polished" (1996, p158). Following this, it was important for me to acknowledge the theories informing the analysis process, which was in part conducted through the use of deductive codes. Throughout my analysis process, it was important for me

to be able to see what connections I was making, and to make this process as explicit to myself as possible.

During the analysis phase of the interviews, it was also important for me to be reflexive in my analysis. For example, when analysing my data, I recognised that although some of my participants were experts in AI, it was necessary for me to analyse their data in a way which recognised their assumptions, working practices, and social relations, rather than simply the assessment of their expertise (Van Audenhove and Donders, 2019). As mentioned before, my participants had a variety of skill levels and were subject to a range of influences and there is a risk that thematic analysis can flatten the varied perspectives and voices in the search for a consistent narrative (Braun et al., 2014). Throughout my analysis, I gave care and attention to each interview irrespective of the participant's place in the project. Furthermore, it was important to respect participants who had concerns about what they had said being misrepresented in the write up of my findings. These concerns were typically due to organisational or project dynamics, where participants were concerned that if others in the organisation were aware of their viewpoints, this could cause the participant difficulty in their place of work. To work through the former issue, I was careful to ensure that views from these participants were fairly represented. To work through the latter issue, I wrote up my findings with privacy at the forefront of my mind. Occasionally, these issues were in tension with each other – to spotlight what felt like the most relevant and important point made by a participant in a junior position in the organisation could mean reporting something said which may cause them difficulty at work if others in their working circle became aware that they had said it. On a couple of occasions, a comment made was difficult to report due to concerns about confidentiality. On these occasions, I would look for similar comments in my data, to avoid potentially highlighting someone with concerns about identification.

Once I had created and reviewed initial codes, I turned to the process of generating themes. This was done by carefully looking through my codes and identifying where these could be grouped. In some cases, I would print out my interview transcripts, cut them into pieces, and physically put them into groups. In others, I would write paragraphs based on particularly prevalent codes, and then look at arranging these in a blank word file, to see how these issues might be related to the theoretical perspectives which influenced my research. After iterating this process, I had something which looked closer to findings. Due to my on-going collaboration with DWP, some topics which emerged from my fieldwork were not analysed and presented as part of my thesis. I expand on this below.

3.9. Additional themes

Throughout my analysis process, I considered which themes were the most saliant to discuss in each research paper. Due to the on-going partnership with the DWP required me to assess which themes I should analyse and discuss in my thesis. This was both due to the sensitive character of the research I was conducting, as well as DWP's influence on the types of theoretical frameworks which would be beneficial for me to use. Due to this, there is a gap between the topics which emerged in the fieldwork due to the character of the collaboration with the DWP and the themes which were analysed and are presented in the following chapters of my thesis. These included themes surrounding dystopian and utopian imaginaries, and neoliberal political ideologies. I further discuss the decisions I made around these themes below.

Regarding the theme of 'dystopian and utopian imaginaries,' this theme was developed within the context of paper two, however the theme was present throughout each of the research papers. From my findings, it was quite clear how dystopian and science fiction had influenced many of my participant's' understanding of algorithmic technologies. Even when participants were arguing for the use of algorithmic technologies, it was often done so through the use of science fiction narratives as a touchstone for communication and creating an idea about the intention of these technologies. These findings were included in paper two (for further discussion, see Chapter 5). Whilst this theme could still be seen in papers one and three, most of what could be said about this theme is already expressed in paper two, albeit limited by the data collected for that project.

Lastly, during the course of each of my research papers I developed themes regarding neoliberal ideologies, particularly in relation to the potential for worldviews to become embedded within algorithmic technologies. This theme was discarded in each of the research papers for two reasons. One, after experimentation, the research papers worked better reporting more on the mechanisms of how worldviews become embedded within algorithmic technologies, rather than the presence and influence of one particular ideology. Two, focusing on neoliberal ideology would have broadened the scope of my research beyond the practical character of my research questions and the DWP context.

3.10. Ethical considerations

Ethics is not only important to research projects to ensure due care is given to participants, but also necessary to ensure epistemological soundness, and that the researcher fulfils their responsibility to society (Sin, 2005; Humphries, 2000; Social Research Association: Ethical Guidelines, 2003). Without ethical principles, a social researcher cannot develop the trust and respectful dialogue between researcher and participant which is necessary to generate authentic findings (Humphries, 2000). Prior to collecting data, ethical considerations were discussed with my contacts at DWP and my supervisors. Following these discussions, I wrote an ethics application for each project and submitted it to my department's research ethics committee. This was an iterative process, in which I would receive feedback on my ethics plan and then resubmit it to the ethics committee to ensure my research design and data

processing plans met the University of Sheffield's ethical standard. During the research process, I paid careful attention to the following issues: informed consent, privacy and ownership of the data, and general sensitivity.

3.9.1 Informed consent

Informed consent is a key component of modern-day ethical research (Sin, 2005). Informed consent was obtained from participants by providing them with information sheets detailing the purpose of the project, and then asking them to sign a consent form stating they had understood and consented to participate in the research. Participants were also given the opportunity to ask questions prior to the interview process or workshop.

While it is traditionally understood that my participants are adults and therefore can freely consent, it was important to consider whether my participants may have had freedom of consent in the context of their work life. As Sin points out, "what constitutes the ability to provide valid informed consent is clearly underlaid by complex ideologies and social constructions of what 'normal', 'competent' and 'informed consent' constitute" (2005, p280). As my research was being supported by the DWP, my participants may have felt pressured into consenting. To attempt to counter this, I ensured participants were aware that they did not have to take part. Additionally, the information sheet informed them that if they opted not to take part, this information would not be passed on to other data science team colleagues or senior members of DWP. However, due to the small numbers within the data science team, potential participants' colleagues may have been aware of their decision not to participate, without it being explicitly shared. While circumstances presented me from fully ensuring confidentiality regarding participants' consent to take part in interviews, I was restricted from fully ensuring confidentiality about consent because of the research context.

3.9.2 Anonymity and ownership of data

In addition, it was important to consider my participants' right to anonymity and ownership of their data during the research process. All transcripts were anonymised, and the original audio files kept on The University of Sheffield's secure Google drive.

After writing paper one, I produced a report to DWP about the findings of the study. In view of the small number of participants in the study, there were concerns that colleagues may be able to identify participants due to their expertise or contextual details. To address this, the report did not include any participant quotations. It was also developed in close communication with DWP and participants (British Sociological Association, 2018).

During paper three, I received a subject access request for a participants' interview video, as per the GDPR and in respect for my participants' ownership of their data, this was provided promptly.

3.9.3 Sensitivity in research

During my research design process, it was important to consider how my participants might feel discussing issues such as discrimination or bias. For example, in my first paper, there were concerns from the University of Sheffield ethical approval committee that participants might feel as though I was looking for someone to blame for biases in their projects. It was therefore important that participants were assured in both the information sheet, and the pre-project presentation, that this was not the purpose of the study.

Other ethical considerations included ensuring participants had contacts in case any of the material in the workshops upset them, due to the workshops' focus on discrimination. Participants were given a DWP phone number they could call in case of any distress. Additionally, participants were provided with both a DWP and University of Sheffield safeguarding contacts' details, in case they had any concerns while participating in the project.

3.11. Limitations

Due to the outbreak of Covid-19 during paper one, the findings presented in this paper may be specific to the Covid-19 context. I am currently unable to assess this, as the full effects of Covid-19 will not be known for many years, and while I have tried to consider how the experience of Covid-19 may have impacted my results, it may have affected them in ways not yet knowable.

Each of the research projects had a sample size suitable to qualitative methods. While the findings on each project can provide insight into the dynamics at play in these situations, they cannot be generalised to other organisations or algorithmic projects.

Particularly in paper two, it is worth noting that my analysis was limited by my restricted knowledge of the Finnish context. During the research process, I worked with a translator to try to understand some of the cultural nuances at play, however due to unforeseen circumstances he was not able to work with me on the project long-term. Before he left the project, he translated three podcast episodes, and wrote a page on how the Finnish words for fairness, ethics, and justice are used within the Finnish language (I asked him to do this to find out if these words had connotations in Finnish which are similar to those in English). I have since spoken to Finnish academics, to try to understand what cultural norms are at play in my data, however as someone who has not been integrated into the culture there are limits to how much I can understand. Furthermore, my participants were unable to use Finnish during the interviews because I am not a Finnish speaker, which in the case of one interview, at times made communication challenging.

During all three papers, it is important to note the possibility of selection biases. As participants volunteered their time to participate in an interview, it is possible my participants were more concerned about algorithmic bias than others working in similar roles. This suggests my findings present a view from those who are somewhat concerned about fixing the issue.

In addition to these issues, much of my research data was generated through interviewing. While I was careful to analyse my data reflexively, it is noted that interviewer effects will influence the type of data generated during this process (Gilbert, 2011, p257).

3.12. Reflection on academic-industry partnerships and the three-paper format

The project brief, three-paper model, and collaborative character of the PhD influenced my approach and decision-making during the course of my PhD research, in addition to influencing what could or could not be said in the process of writing up my thesis. I needed to maintain good working relations with my contacts at DWP and participants, handle concerns regarding individual and organisational confidentiality, and navigate data collection practices for a three-paper thesis whilst managing the longer-term goals of the thesis. I expand on these below.

Due to the collaborative nature of my research, it was important I maintained good working relations with my contacts at DWP. DWP were supportive and engaged throughout my PhD research, and we worked together to come to an agreement about how to approach each research paper. Each project was designed in a way which aligned with the DWP working culture and the working practices of its employees. For example, because of concern at the DWP regarding media reports of their working practices, it was important I respected DWP's boundaries regarding what could and could not be said publicly about the data science projects I studied. Furthermore, these concerns shaped which projects DWP facilitated my access to during my research. Whilst the area DWP are most reported to use algorithmic technologies in is the Fraud Detection department, this area of the organisation is considered especially sensitive, and was deemed not suitable as the focus of my research. Furthermore, the study of fraud algorithms presents its own challenges, including the release of sensitive information which might enable some people to 'game' the system, a concern found by Veale et al. (2018) in their research with public sector workers in a US context.

The partnership relationship with DWP influenced the approach I took to analysing and presenting my research, particularly due to the current character of social science literature relating to AI and algorithmic developments. As it stands, much of the literature is critical of the deployment of algorithmic technologies, for reasons which I am sympathetic to. However, given my partnership with the DWP, I needed to ensure my research would be useful to those working in the organisation. During the PhD process, it was vital for me to write in a way which respected the culture, values, worldviews and frameworks used by the organisation and the efforts made by the people working there, whilst also conveying potentially challenging insights. At times, this was challenging, due to my own political leanings and values regarding the progression of social justice. However, it was important for me to adopt a framework and style of writing which conveyed both my participants' beliefs, whilst not compromising my own values and integrity as a researcher.

Furthermore, this process gave me a lot of experience thinking about how social scientists might balance factors such as research integrity with producing research which organisation's find actionable. As I considered myself to be conducting applied social science research (for more information on applied social science research, see Chapter 3), a framework which is rooted in the expectation of tangible research outputs, it was important for me to consider how my research might be received within the organisation. Factors on which it was important for me to reflect on throughout my research process included suitable language for conveying challenging information, and the types of insights an organisation such as the DWP might be able to action. At the same time, I had a responsibility to report my research findings in a way which did not leave out nor sideline material which may have been challenging to DWP.

Furthermore, the three-paper model presented its own academic challenges. Prior to starting on my PhD, I had only completed one research project as part of my masters. The format of the PhD required me to design and execute three research projects, one for each year of the PhD. Additionally, the structure of the thesis would mean that each of the empirical chapters of my thesis would need to meet the expectations of an academic journal paper. This presented two challenges, identifying themes and narratives which would be suitable to be presented in a paper-based format, and learning to write in a manner which suits the journal paper genre early in the PhD process. For example, after initially analysing and writing up the findings on each research paper, I often found the paper was far too long for the journal paper format. I had to be careful about which themes I reported on, so each paper had an uncluttered narrative which was suitable for a journal paper. For further discussion of themes that emerged but were not included in the thesis, see Chapter 3.

Additionally, the working relationship with DWP and the paper-based format affected data collection and what data might be available for me to analyse in my thesis. For example, due to the paper-based format I was required to apply for ethical approval prior to each individual research project and stipulate my data collection methods within the context of an individual paper. However, my understanding regarding DWP working practices developed from meetings with my contacts and the internship I completed with them as part of the masters' components of my PhD. This meant some elements of my understanding of DWP did not emerge from research data. Whilst some of the information that led to my understanding was repeated during interviews with participants, it was not possible for all these insights to be included in my research data, because they did not emerge from the data gather under ethical approval.

Whilst this experience bought with it various challenges, this experience was particularly valuable in learning how applied research and research impact might be navigated during the research process.

3.13. Conclusion

This chapter has detailed my methodological approach to my PhD thesis. The chapter started by discussing how the 3-paper model influenced my design choices, my high-level overarching research aims, and how these were turned into research questions for each research project. After this, I discussed the DWP context and details relevant to interpreting my design decisions, findings, and analysis. I then discussed my research design process, including the selection of suitable projects, and how I created the criteria for project selection. Following from this, I discussed my data collection methods, and how these were employed on each research project. Then, I discussed my data analysis process, ethical considerations, and the limitations of my research. This ends the introductory section of my thesis. The following chapters will be the three empirical chapters which form the basis of my thesis.

Chapter 4: Paper one, Investigating the role of current DWP working practices in mitigating algorithmic bias

In the previous chapters, I described an increase in the use of algorithmic technologies within the public sector, accompanied by increasing cases of the phenomenon known as algorithmic bias. In response to this, the UK civil service organisation the DWP supported my PhD to gain a better understanding of issues of algorithmic bias and how algorithmic bias may be mitigated. After reflecting on the literature and my research problem, I devised the following thesis structure:

- *paper one:* investigate what DWP data scientists are currently doing in areas related to algorithmic bias
- *paper two:* investigate how other organisations are successfully attempting to mitigate algorithmic bias
- *paper three:* investigate how could the insights from paper two be integrated into a DWP context

In this chapter, the first of these papers is presented.

4.1. Introduction to paper one

Industry experts have proposed that algorithmic technologies will play a major part in the next industrial revolution (Daws, 2020). In turn, organisations have rushed to transform their working practices in various ways: through integrating data-driven methods; by collecting and repurposing data to provide organisational insights; and by using algorithmic systems to support staff decision making. This data-driven turn focuses almost exclusively on quantitative tools – purportedly allowing for a better understanding of staff, customers, organisations, and the world at large. Typically, algorithmic technologies have been considered to belong to the specialism of data science, where these technologies are conceptualised as a form of narrow AI (artificial intelligence). Narrow AI refers to an automated model limited to performing a specific task, for example recognising and recording the license plate numbers found by speed cameras (Bostrom, 2014) (for further discussion, see Chapter 2).

Against this backdrop, public sector services have begun using algorithmic technologies to take advantage of their supposed efficiency gains. This builds on these organisations' previous work using statistical modelling, in which practitioners would use quantitative techniques to provide empirical insights in the development of services and policy (Monaghan and Ingold, 2019). UK public sector organisations that have started using, or expressed interest in using, algorithmic technologies include the DWP (Department of Work and Pensions), local authorities, police constabularies and other civil service organisations (Oswald *et al.*, 2018; Trendall, 2019). In addition, in 2021 the UK government released its

National AI strategy, expressing the importance of AI in promoting resilience, productivity, growth and innovation in the public sector (Office for Artificial Intelligence, 2021).

As noted in previous chapters, the adoption of algorithmic technologies in the public sector has not been without consequence (Angwin *et al.*, 2016; Eubanks, 2018; Noble, 2018) . As the concerns surrounding these technologies have grown, organisations have released guidance documentation to assist data workers in managing the risk of algorithmic bias. Among these are documents published such as the Central Digital and Data Office commissioned Algorithmic Bias Landscape Summary, a document which aims to draw together the literature on algorithmic bias, as well as potential strategies and methods to mitigate against it (Rovatsos, Mittelstadt and Koene, 2019). Furthermore, influential consulting agencies have released their own AI guidance documentation, such as PricewaterhouseCoopers's responsible AI toolkit (PricewaterhouseCoopers, 2019). Data science practitioners have also contributed, publishing academic papers designed to provide practical guidance for mitigating algorithmic bias (Cramer *et al.*, 2018). These publications informed my research on this paper.

However, while there has been a surge in algorithmic bias mitigation guidance, there has been little research about how data scientists currently engage with the themes identified in these guidance documents. This paper contributes towards understanding how practitioners engage with the details of guidance documents. To investigate this, I conducted eight interviews with six civil service practitioners at the DWP, concentrating on an active DWP data science project, and examining how staff engage with themes identified in three separate guidance documents. In this paper, I present two findings. The first finding I present is that my participants strongly relied on legal frameworks due to their position as civil servants; however, these legal frameworks did not facilitate accountability to the population they served. Secondly, my participants bias checking practices were limited by previous research conducted by the DWP, in addition to influences from the department's organisational culture.

This paper will proceed as follows. First, a literature review focusing on the current guidance surrounding algorithmic bias mitigation, specifically technical and legal guidance, and the effectiveness of guidance documents more generally. Second, a methodology section detailing my research design and a description of the DWP data science project investigated. In the third section, I discuss my findings in relation to the following themes: 'reliance on legal frameworks', which focuses on how practitioners in a DWP context engage with legal frameworks within their working practices; 'collaboration through documentation', which focuses on how project documents are used to facilitate inter-team collaboration; 'responsibility and accountability', which focuses on my participants' perception of their responsibilities and accountabilities in their working practices; 'bias checking', which discusses the types of bias checking my participants were actively engaged in; and 'the influence of organisational culture', which focuses on how organisational culture led by government practice is a contributing factor in the development of biases. Finally, I discuss the implications of these findings.

4.2. Literature review

Guidance for mitigating bias within algorithmic systems

As actors attempt to mitigate algorithmic bias, a number of guidance documents (such as voluntary standards, codes of practice, and ethical frameworks) have been produced in an attempt to address this issue (Kuziemski and Misuraca, 2020). However, the role these documents play in mitigating algorithmic bias is still not well understood. Several organisations have produced guidance documentation for this purpose (Dencik *et al.*, 2018, p118), with documents being produced by government offices, researchers, consulting agencies, and professional bodies (Cramer *et al.*, 2018; PricewaterhouseCoopers, 2019; Rovatsos, Mittelstadt and Koene, 2019; *Data Ethics Framework*, no date). Often, the contents of these documents focus on which laws and regulatory guidance practitioners are required to follow, who is accountable for ensuring algorithmic bias mitigation practices have been followed, and the technical tests practitioners can perform to review biases within their datasets (Jobin, Ienca and Vayena, 2019). However, scholars within critical algorithm studies have expressed concerns that the advice offered by these documents provides neither adequate understanding of the mechanisms of algorithmic bias, nor an effective pathway for mitigating the risks posed. I expand on these critiques below.

Technical guidance

Guidance documents often suggest practitioners use technical tests to investigate issues of bias, for example using comparative statistics to check for disparities between different demographic groups (Cramer *et al.*, 2018; Jobin, Ienca and Vayena, 2019; Rovatsos, Mittelstadt and Koene, 2019; *Data Ethics Framework*, no date). These types of technical approaches use well established techniques in the fields of data science, statistics, and AI development. However, critical data scholars have challenged these approaches (for a more in-depth discussion of these arguments, see Chapter 2: FATE approaches). Specifically, academics have been critical of these types of technical tests, due to concerns that they further a narrow conception of fairness (Green and Hu, 2018).

As discussed in Chapter 2, Green and Hu (2018) identify two key means by which fairness has been operationalized by data practitioners when investigating bias using technical approaches; statistical fairness tests, and procedural fairness methods. Statistical fairness, also known as parity-based fairness, is when statistical outputs such as accuracy scores and comparative statistics are generated so that outputs from different groups can be compared with each other. Following from this comparison, practitioners can judge the fairness of a model based on how balanced these appear (Green and Hu, 2018, p2)**.** For example, if practitioners were developing a fraud detection algorithm, they might compare the number of false positives and false negatives produced across all demographic groups, to ensure no group is receiving a disproportionate number of false positives or negatives. In contrast, procedural fairness is concerned with ensuring the processes involved in data collection, data

analysis, and the evaluations made during development are the same across demographic groups (Rovatsos, Mittelstadt and Koene, 2019, p11). For example, if someone were to develop a hiring algorithm, it would apply the same steps, procedures, and expectations to each applicant.

Green and Hu argue that both these methods "capture important considerations of fairness: impartiality of process on the one hand and protection from adverse impact on the other" (Green and Hu, 2018, p2). However, critics, including Green and Hu themselves, argue that concentrating on these two aspects of fairness is not enough to mitigate the risks of algorithmic bias (Green & Hu, 2018; Hoffmann, 2019). The argument is that these processes enforce a narrow and ahistorical understanding of fairness, and that by turning away from the broader context of injustice "we run the risk of overlooking systemic issues and deeming social structures fair simply because we have improved one component of them" (Green & Hu, 2018, p4; Hoffmann, 2019).

Legal guidance

Furthermore, as organisations shift more of their human decision making to algorithmic systems, the question emerges of what organisations' responsibilities are regarding the decisions made by algorithmic technologies. In 2018, the GDPR was implemented across the EU, which included various mechanisms pertaining to algorithmic bias, particularly those relating to the responsibilities of data controllers (GDPR, 2016). Whilst the UK withdrew from the EU in 2016, the GDPR still applies within a UK context. Specifically, article 22 outlines the data subjects' 'right to an explanation' as to the logics of any fully automated decisions made about them (GDPR, 2016). As noted in Chapter 2, Edwards and Veale (2018) argue that citizens are often unable to exercise the 'right to an explanation,' due to the inaccessibility of these mechanisms. Edwards and Veale put forward this argument using the example of data protection Subject Access Requests (SARs), which they argue have been predominantly used by journalists and company insiders, and not the public at large, due to the amount of time, knowledge, and persistence that following through with these requests requires (p6). Furthermore, if a data subject were to receive an explanation as to how an algorithmic system came to its decision, it is uncertain how this might assist the data subject in receiving a fairer outcome (*ibid*).

Additionally, article 35 stipulates that if the data controller utilises new technologies to process data, and the type of data processing is likely to present a high level of risk to the rights of data subjects, then a data protection impact assessment (DPIA) must be undertaken prior to data processing (Edwards and Veale, 2017). These legal shifts establish mechanisms by which organisations bear some legal accountability for considering the impact of these technologies, in addition to accountability for documenting those considerations. Leonelli (2016) describes what it means for someone to be accountable, rather than merely responsible, as follows; "responsibility [is] the moral obligation to ensure that a particular task is adequately performed […] accountability denotes the duty to justify a given action to others and be answerable for the results of that action" (Leonelli, 2016). These mechanisms assign

data controllers some degree of accountability for how potential impacts are considered within their organisations.

However, critics have argued this is insufficient, and that while this creates organisational accountability regarding the production of these documents, these documents provide data subjects with very little practical protection against harm (Metcalf, *et al.*, 2021). Moreover, there is no stipulation that those who may be subject to potential impacts of these technologies are consulted in the production of these assessments, prompting concern that the documented impacts will bear little resemblance to the harms faced by the groups of people most likely at risk of harm (Metcalf, Anne Watkins, *et al.*, 2021; Yam and Skorburg, 2021).

In addition, the GDPR specifies data subjects may request that decisions made about them are not solely made on the basis of an automated system, which has informally been referred to as the 'human in the loop' requirement (Wachter, Mittelstadt, Floridi, 2017; GDPR, 2016). However, it has been suggested that these protections may be ineffective in practice, due to the rarity of situations where algorithmic systems are the sole decision makers, thus providing data subjects no protection when human decisions are influenced by an automated system (Edwards and Veale, 2017). This is of particular concern, as research has demonstrated human-decision makers will default to algorithmically generated recommendations in certain contexts (Eubanks, 2018).

While there has been extensive criticism surrounding algorithmic bias and algorithmic bias mitigation guidance, there has been little empirical research as to the role guidance documentation plays in mitigating the risks of algorithmic bias in practice. In the next section, I review the limited literature on the effectiveness of guidance documentation.

Effectiveness of guidance documentation

Algorithmic bias mitigation guidance is often considered a sub-category of ethical AI guidance (Jobin, Ienca and Vayena, 2019). Ethical guidance documents are documents "containing a company's philosophy and rules of ethical and acceptable behavior" (Rorie and West, 2022; Schell-Busey, 2009). However, research on these types of documents has been beset with conceptional difficulties. For instance, empirical research in this area may refer to organisational codes of conduct, ethics documentation, or professional codes of practice, making it difficult to assess whether such research refers to the same type of document (Rorie and West, 2022). Furthermore, researchers rarely include the documents within their methodological materials. This presents challenges in attempting to establish the contents of these documents and the differences between them. To address this conceptual difficulty, I have included literature from a broad range of guidance documents which share the aim of influencing professional and organisational communities to adhere to specific standards concerning moral behaviour.

There are few studies which investigate guidance documentation in a software context. One exception is McNamara *et al.*'s (2018) experimental study, in which software engineering

students and professionals were instructed to answer questions relating to two ethical vignettes. The first of these was similar to the Volkswagen Emissions Scandal, in which some Volkswagen vehicles had software installed which allowed the car to deceive environmental emissions standards tests. The second focused on self-driving cars. Participants were split into two groups; in one, participants were explicitly instructed to answer the questions while considering the ACM code of conduct, which was designed to inspire and guide the ethical standards of all computing professionals (ACM, no date). The other group of participants were not given this instruction. Their study concluded that the instructions had no impact on the software engineers' decisions (McNamara *et al.*, 2018: Green, 2018). However, their findings suggested that individuals in the group who were instructed to read the ACM code of conduct were more likely to follow the codes of conduct if they were aware of the high-profile news stories on which the vignettes were based (*ibid*). The authors interpreted this to mean that individuals were better able to act in accordance with the codes of conduct if they could connect them to high-profile examples.

In a similar study on the decision-making processes in auditing and accounting professionals, professional ethics codes were found to have a positive impact on decision making. However, this effect was only found in participants with greater experience in the role to which the ethics codes were related (Pflugrath, Martinov-Bennie, and Chen, 2007).

Some researchers have gone beyond investigating the effect of guidance documentation specifically, to investigating the culture which exists around the guidance. For example, Bowyink's (1994) qualitative study on the influence of ethical codes in journalistic newsrooms revealed two key findings. One, management's perception of the importance of ethical codes was highly influential – ethical codes were more effective when held in high regard by organisational management. Relatedly, Slaughter *et al.* (2004) found that codes of conduct are better adhered to when they are strongly, rather than weakly enforced. However, this effect was only found in individuals who scored as highly conscientious on a personality test which was administered as part of the study.

Additionally, Bowyink (1994) found that in organisations which viewed their ethical codes as effective, those working in the organisation did not expect them to act as strict guidance, but as a starting point for debates and conversations about the moral issues they encountered within their work (*ibid*). Furthermore, in Schwartz (2004)'s qualitative study at four large Canadian companies, findings suggested the importance of employees' understanding of codes of ethics. Specifically, the study found employees experience codes of ethics as easier to understand if examples are provided, with participants expressing that codes of ethics were not likely to be effective if examples were not present (*ibid*). Moreover, their participants suggested documents needed to provide them with instructions on what not to do, rather than suggest they should try to do something, to be effective. The findings of these studies suggest the importance of organisational context in how guidance documents are perceived, in addition to the need for these documents to effectively communicate expected behaviour. However, there is still uncertainty as to how the organisational context influences the effectiveness of guidance documentation in the context of algorithmic bias, the focus of this study.

Turning to consider the role of guidance documents within a public sector context, Christensen and Lægreid (2011) found that the adoption of ethical guidelines aimed at improving and developing ethical awareness within the Norwegian civil service had a mixed reception. Their study found that civil service workers who held a higher position in the organisational hierarchy viewed ethical guidelines more positively than those who held a lower position (*ibid*). In addition, their findings suggested some ministries adopted ethical guidance more quickly than others (*ibid*). They suggest that this difference resulted from the fact that the guidance had been written in a way which formalised the values already practiced within these ministries, and thus were already supported by the organisational culture and working practices within these departments (*ibid*).

Typically, research has focused on the 'effectiveness' of guidance documentation (Jensen, Sandström and Helin, 2009a), generally measured by adherence to the guidance in question. As presented here, the effectiveness of codes of practice is complicated, with questions arising around what constitutes e*ffective* in this context. The examples here also suggest that in addition to the codes themselves, much of their 'effectiveness' relies on organisational culture and individuals' dispositions, alongside the guidance itself. Following from this, I outline the methodology used to investigate how DWP data science practitioners engage with elements of existing algorithmic bias mitigation documentation.

4.3. Methodology

Research methods

This study addressed the research questions; RQ1a: what algorithmic bias mitigation working practices are currently engaged in by data science practitioners at the DWP? and RQ1b: What are the limitations of these practices? These questions were addressed in three stages. In the first, I utilised desk research to identify the key areas on which algorithmic bias mitigation guidance focuses. In the second, I interviewed practitioners actively working on a DWP data science project, to investigate how they engaged in the areas identified. In the final stage, I transcribed the interviews and analysed them using thematic analysis. A fuller description of this process can be found below.

In the first stage, I sourced bias mitigation documentation and analysed it to understand its content. As noted in Chapter 3 above, to find suitable documents, I searched for currently available algorithmic bias guidance documents in academic, grey, and commercial literature which focused on algorithmic bias mitigation. This search was conducted through a combination of Google Scholar, Web of Science, and internet search sites such as DuckDuckGo and Google. During this search, my criteria for inclusion were: a) the document appeared as though it could be translated into somewhat practical advice by an organisation, b) it appeared applicable to a DWP data science context, and c) was aimed at an organisational level which would be relevant to my participants.

Examples of topics which were considered suitable included suggestions about what data science teams should be doing to avoid algorithmic bias, techniques which may or may not

be favourable, and setting out a plan for future work in an organisational context. Three guidance documents were selected as relevant to the DWP context. These were PwC's responsible AI toolkit, The Centre For Data Ethics and Innovation Landscape Summary on Algorithmic Bias, and Cramer *et al.*'s framework to algorithmic bias (Rao, Palaci and Chow, 2019; Rovatsos, Mittelstadt and Koene, 2019; Cramer *et al.*, 2018) – that is, the three documents listed in the introduction above.

The PwC responsible AI toolkit provides high-level management advice around developments in AI, focusing on fairness, governance, security, and organisational workflows (PricewaterhouseCoopers, 2019). The Centre For Data Ethics and Innovation Landscape Summary draws together academic literature and public sector concerns to provide an overview of algorithmic bias and some potential solutions (Rovatsos, Mittelstadt and Koene, 2019). The mitigation methods proposed included technical bias checking, procedural fairness methods, and project management. Finally, the Cramer *et al.* (2018) framework on algorithmic bias explores mitigating bias within the context of an organisation in the music industry. This focuses on how to assess algorithmic bias in practice, and how to communicate with other teams.

Once I had selected these documents, I carefully analysed them for themes to inform my interview guide. During this process, I identified the following themes: strategic aims of the project, project management, regulations, project documentation, and model decisions. Following this, these themes were used to develop my interview guide. Additionally, the themes developed at this stage were used as codes during the data analysis, to guide analysis of the guidance documents.

In the second stage of the project, I interviewed practitioners about these areas in relation to a DWP project they had recently worked on. At the time of data collection, the Sheffield DWP data science team had been working on two projects. As one of the projects was still in the design stage, the most suitable project for the interviews to focus on was the other project, the Digital Trialling Framework (DTF), and the associated Digital Plus Trial. This was a project jointly run by the data science team and a specialist policy analyst team at the DWP, called the Labour Market Policy Group. A fuller description of these projects can be found in section 4.2.1. below.

The semi-structured interview approach was chosen because it is an appropriate research method to understand individuals' experiences and the social processes in which they engage (Delmont and Mason, 1997; Bryman, 2012; Van Audenhove and Donders, 2019). Additionally, interviews with professional experts can provide rich data about a) the bureaucratic and working practices which surround the spaces where algorithmic bias might develop, and b) how the context, power, and structure within the organisation may stifle – or enhance – efforts at solving societal conflict (Van Audenhove and Donders, 2019). See section 3.7.1 and section 3.7.5 for further discussion.

For this study, I interviewed six participants, with eight interviews conducted overall. The sampling criteria were that the individuals had worked on either the DTF project or Digital Plus Trial, in addition to having worked at the DWP for at least six months, to ensure participants were sufficiently embedded within DWP organisational culture to meaningfully engage with the questions posed. Out of the six participants, four were from the Data Science team at the DWP, and two were policy analysts from the Labour Market Policy Group (a group which trials potential policy or service interventions with the aim of bringing people closer to the labour market). Two participants from the data science team were interviewed twice; first at the beginning of the study, and second at the end of the study, to gain their perspectives once the project had finished.

The data collection period for this paper fell between February 2020 and August 2020. Interviews lasted between 40 and 120 minutes. Due to the outbreak of Covid-19, the context of the interviews changed across the data collection period. Additionally, both the DTF project and Digital Plus Trial were terminated during this time period due to Covid-19. As a result of this interruption, participants sometimes found it difficult to remember project details in later interviews. However, one participant mentioned that while he was less able to remember specific details, he felt able to speak more freely about what could have been changed on the project due to the project's termination. These details, although not planned, still provided a rich understanding of participants' experiences on the DTF project and Digital Plus Trial project.

After conducting the interviews and data collection, audio files of the interviews were transcribed into text format. Once transcribed, the interviews were analysed using thematic analysis (Braun *et al.*, 2014). This included use of both inductive and deductive coding. Inductive codes were built around themes that had appeared in the guidance, and were developed depending on the themes found within that cluster. For further details of my analysis process, see Chapter 3.

This study underwent ethical approval at the University of Sheffield, to ensure that the topics and the research design did not endanger participants and were sensitive to the context of the organisation. Specifically, the ethics application for this paper paid careful attention to ensuring participants were aware that I was not investigating the project to apportion blame for biases or non-adherence to policy, but to generate a fuller understanding of their working practices, to form the basis of my understanding for subsequent projects. I provide a fuller description of the DTF project and Digital Plus Trial in the following section.

Description of Project

The data collected for this study centred around the DTF (Digital Trialling Framework) project and the Digital Plus Trial. These were two separate projects developed concurrently to provide the department with the means to carry out real time randomised control trials (RCTs). RCTs are an experimental quantitative method which is meant to identify causal

mechanisms through the use of a control and treatment group. The Digital Trialling Framework involved the development of a 'data pipeline' to allow DWP policy analysts to access data from the new Universal Credit (UC) system for use in policy and service intervention trials.

As identified in Chapter 3 (Section 3.5: The DWP context) above, Universal Credit is a DWP social security benefit service which combines and replaces six older benefits, which include; income-related Employment and Support Allowance, income-based Jobseeker's Allowance, Income Support; Child Tax Credit and Working Tax Credit; and Housing Benefit (GOV.UK, 2020). Furthermore, the UC system is an 'all-digital' service. This means claimants are now required to register for the service online, in addition to providing digital evidence of their engagement with the DWP service. This shift to digital has created a new source of data within the organisation, and during my interviews, participants explained that the department were investigating how to make better use of this data. The DTF 'data pipeline' was being designed to assist the department in achieving these aims, allowing policy analysts near real time access to the UC database. As one participant explained:

> "We were kind of in a new space as well with Universal Credit, in terms of the amount of information that we can get [from] the system, [its] huge - there's masses of it […] not all of that we would need, [so] the idea was to kind of find a way of developing... what I think was called a pipeline [to filter that data]" (P5, Analyst)

Furthermore, the DTF 'data pipeline' intended to increase the speed policy analysts were able to provide ministers with results, facilitated by the development of a visual dashboard so the data could be easily understood. One participant explained:

> "we were kind of required or challenged by the relevant minister that we could turn the findings round much more quickly […] ministers and senior officials, understandably they want to know how its working, and they want to get information quickly" (P5, Analyst)

The Digital Plus Trial was designed as the pilot trial for the DTF pipeline, with the Digital Plus Trial component of the project managed by the Labour Market Policy Group. The aims of the trial were to test the hypothesis that Jobcentres could improve their services if some job seeking claimants were able to engage with the department digitally, rather than attending face-to-face sessions with their work coach. As discussed in Chapter 2 (Literature review), DWP policy has progressively developed to include more 'conditionality', meaning the claimant is expected to meet a growing number of conditions to continue receiving their UC payments. These conditions involve completing 35 hours of work search activity including searching for jobs, applying for jobs, and attending training courses and interviews. Additionally, each week the claimant attends a meeting with their work coach at the Jobcentre to discuss their week's work search activities. If claimants do not complete these

tasks, they may receive a 'sanction', meaning a claimant will not be paid for a set amount of time. Claimants may also receive a sanction for not attending their work coach appointments.

The Digital Plus Trial was being designed to investigate whether work coaches could save time by engaging with the more 'digitally competent' claimants with 'less complex needs' online, instead of seeing them in a face-to-face setting. It should be noted this was pre-Covid-19, when meetings typically occurred in face-to-face settings. Furthermore, the trial aimed to provide work coaches with more time to spend with claimants with 'complex needs', a category which included claimants experiencing domestic violence, involved in gangs, with specific disabilities, or without the digital skills necessary to partake in the trial (GOV.UK, 2022).

The Digital Plus Trial was piloted at 26 Job Centres. The data science team worked through the technical challenges of retrieving reusable data from the UC database, producing the monitoring dashboard, and developing a system to track the different cohorts of the trial. The labour market team colleagues focused on the analytical difficulties presented by the trial design itself.

Both the DTF project and the Digital Plus Trial were cancelled after the outbreak of Covid-19, and the Digital Plus Trial and DTF projects did not go beyond the proof-of-concept stage. While the projects were not integrated into the DWP's standard organisational practices, the learning generated from these projects was taken forward as the DWP moved to more online processes post-Covid-19. Furthermore, the project is a good example of the types of data projects DWP data science teams develop.

4.4. Findings

The findings below are organised around three themes. The first is 'reliance on legal frameworks', focusing on how practitioners in a DWP context engage with legal frameworks within their working practices. The second, 'collaboration through documentation', focuses on how project documents are used to facilitate inter-team collaboration. The third, 'bias checking', discusses the types of bias checking my participants were actively engaged in. The fourth, 'considerations about how biases emerge', focuses on how government practice is a contributing factor in the development of biases. And finally, 'the influence of organisational culture', focuses on how organisational culture led by government practice is a contributing factor in the development of biases.

Reliance on legal frameworks

Algorithmic bias mitigation guidance often reinforces the importance of data practitioners' adherence to legal frameworks (Jobin, Ienca and Vayena, 2019; Rovatsos, Mittelstadt and Koene, 2019; *Data Ethics Framework*, nd). This section explores how these practitioners engaged with legal frameworks, and practitioners' perceptions of the importance of these legal frameworks within their work.

My participants expressed the importance of their understanding of regulatory frameworks to ensure both the DTF project and Digital Plus Trial were conducted in a way which was ethical and free of bias. Often, when participants used the term 'ethical', it was in the context of following legal guidelines. For example, one data science participant said:

> "depends on how the governance will catch up, because that's always going to be our hurdle, you know there's always a lot of people in the middle who won't do things because […] we haven't thought about those things, […] but this is a good thing right, because it makes sure we are being ethically responsible, technologically responsible, and we do things properly […] from a GDPR type of view" (P2, Data scientist)

Participants described the standard procedure each project had to go through prior to receiving clearance. During the initial stages of the project, the data science team completed an application form which detailed their reasons for needing the data in question, why it was needed for the department to fulfil the mandate given to it by the government, and how the project would benefit DWP service users. Once the application had been completed, and the member of the team responsible for monitoring GDPR had assessed the application, the form was passed on to a specialist team to assess whether the project and data request was appropriate for the organisation. Participants described the importance of considering issues such as data minimisation during this process, so only requesting as much data as would be needed to fulfil the aims of the project. Furthermore, the application process had a strong focus on data safety and security, and any application would need to demonstrate understanding of the appropriate manner in which sensitive data should be handled. Once the specialist team assessed the application, the data science team was informed of the outcome.

One participant described this process as being lengthy, saying that it could take up to six months from the time of requesting the data to receiving the data. There were two reasons for this. First, the team which assessed the application might take a long time to read through and decide the outcome of the application. Second, once the data had been authorised, the servers would need to be updated to allow access – a process which occurs at specific points during the year. If authorisation was provided after this date, the team had to wait until the next time the servers are updated before they could receive the requested data.

The regulatory framework my participants primarily referred to was the GDPR. Furthermore, they explained that the department had numerous resources which supported them in both understanding and fulfilling the mandates passed to them through these legal structures. These resources included specialist training for all data handling staff, incorporating specialist staff members who were responsible for ensuring compliance in the team, and bureaucratic processes designed to ensure all the required legal steps had been taken. Most participants exhibited high levels of awareness regarding these legal structures. Moreover, they conveyed a sense of respect and appreciation towards the perceived thoroughness and protection their

training and the organisation's approach provided them in their work. One participant commented:

> "It's not for us to decide what we do because we are here employed by the department so each project goes through the [standardised process] […] our admin manager goes through all of the project's mandate, [where] we describe the project, what we're going to do, why it's needed, what's in it for our customers including how we can do it ethically under GDPR and covering all the bases" (P1, Data scientist)

Participants described the standard procedure each project had to go through prior to receiving clearance. During the initial stages of the project, the data science team would need to complete an application form which detailed their reasons for needing to use the data in question, why this data was needed for the department to fulfil the mandate given to it by the government, and how the project would benefit DWP service users. Once the application had been completed, and the member of the team responsible for monitoring GDPR had assessed the application, the form would be passed on to a specialist team to assess whether the project and data request was appropriate for the organisation. Participants described the importance of considering issues such as data minimisation during this process, so only requesting as much data as would be needed to fulfil the aims of the project. Furthermore, the application process had a strong focus on data safety and security, and any application would need to demonstrate understanding of the appropriate manner which sensitive data should be handled. Once the specialist team assessed the application, the data science team would be informed of the outcome.

One participant described this process as being particularly lengthy, saying that it could take up to six months from the time of requesting the data to receiving the data. This was due to two reasons. One, the team who assessed the application might take a long time to read through and decide the outcome of the application. Two, once the data had been authorised, the servers would need to be updated to allow access – a process which occurs at regular intervals during the year. If authorisation was provided after this date, then the team would need to wait until the next time the servers are updated before they could receive the requested data.

As found in Orr and Davies' (2020) study, participants relied on legal codifications to assess their ethical responsibilities, which Orr and Davies describe as practitioners off-loading responsibility for ethical debates onto adherence to legal codes (Orr and Davis, 2020). In their study, Orr and Davies interviewed practitioners from the private, public, and academic sectors. However, their findings did not specify to which sector these findings related. My findings suggest a more complicated picture. While participants highlighted the strong organisational culture built around legal compliance, they did not talk about this as though its purpose was to offload legal responsibility, but instead emphasised their duty of care to citizens, and their obligations as civil servants. Rather than trying to do the minimum,

participants saw it as carrying out the responsibilities of the government to the public. Or, to put it another way, participants thought it was not their responsibility to decide what the right course of action would be, due to their duty as civil servants to adhere to established procedures. This permeated through much of the practice my participants described. For example, in discussing standard DWP procedures on a data science project, one participant said:

> "Because we are in DWP we have data to fulfil our duties for the public, we need to have some data […] so it's not because we are interested, it's because we need it, so it's the minimum data we need to collect to fulfil our duty and make their lives easier, and we help them when they are in their difficult moments." (P1, Data scientist)

Even when discussing regulatory practices such as data minimisation principles, where data controllers must only use the minimum data required to perform the task in question, this participant refers to the moral reasoning ("to fulfil our duty and make their lives easier") to justify why the data is being processed, in relation to DWP's duty of care to citizens. Furthermore, as found in Veale *et al.*'s (2018) study of US public sector data workers, participants did not use sensitive information in their models, such as sex, race, or postcode, unless the situation warranted this data.

Collaboration through documentation

Additionally, participants referred to the importance of completing a DPIA (Data Protection Impact Assessment) and engaging with a data security colleague before data could be released for use by the data science team. Metcalf *et al.* (2021) suggest impact assessments can act as boundary objects in algorithmic bias mitigation. Boundary objects are objects such as documents, schemas, concepts, etc., which have "specific meanings for experts within disciplines, but are malleable enough to hold their meaning across disciplines and become productive sites of collaboration" (Star, 1989; Metcalf, Moss, *et al.*, 2021) – for example, a standardised form used in a hospital to record patient symptoms (Star, 1989). Such a form can be filled in by different types of clinical practitioners, and may be used by researchers in clinical studies (*ibid*). These documents are understood locally within disciplines, but are malleable enough to be understood across contexts (*ibid*). Metcalf *et al.* (2021) suggest impact assessments provide experts with the means to collaborate and co-construct their understanding of potential impacts (Metcalf, *et al.*, 2021).

Through an impact assessment, experts from different disciplines are able to imagine, communicate about, and decide upon the range of potential impacts which may arise from the data science project in development. During the development of the DTF and Digital Plus Trial project, both Labour Market Group and Data Science team members explained how the project required them to learn to collaborate and learn about each other's working practices, as they had not previously worked together. In addition, participants explained that the Labour Market Policy Group was the Data Science team's client on the project. As such, overall responsibility for the project rested with the Labour Market Policy Group Although both teams had a strong basis in statistical analysis, participants described that it was nonetheless

important for each group to assess and decide upon their remit during the project. Talking about how the teams learned to work together, one data science participant described the issue the teams had working together at the beginning of the project, saying:

> "'cause we're not policy analysts, we're data scientists, right […] I think that's where the early project team went a little bit wrong, because they were trying to comment on the trial design, and that is just not our speciality" (P2, Data scientist)

Participants from both teams described how they manoeuvred in order not to encroach upon the other team's speciality, expressed in the form of respecting each team's expertise. The beginning of the project involved each team identifying their remit in the project. This occurred not only regarding the trial design, but also in terms of what assessments each team was responsible for. While the data science team had been engaged with the data protection impact assessment, which focused on why there was a need for certain types of data, they had been less involved in assessments which were outside of their 'remit', such as the Digital Plus Trial's ethics application.

The Digital Plus Trial's ethics application involved the trial team and the DWP's ethics committee assessing the trial's potential for harm to claimants or DWP workers. However, the ethics application process was considered outside of the data science team's remit. When participants discussed the ethics application for the Digital Plus Trial, they explained that it had been completed through typical DWP trial procedures. While the Labour Market Policy Group participants had been involved in the ethics application for the digital trial, members of the data science team were not entirely sure if they had seen some of the initial ethics documentation. In this case, the impact assessment and ethics application did not function as boundary objects to facilitate communication on the potential harms of the project, due to participants' desire to stay within their own realm of expertise to ensure the project ran smoothly.

Responsibility and accountability

Additionally, it is worth noting that these documents did not engage those who would be at risk of being impacted by the new technology – the claimants themselves. Metcalf *et al.* state that engaging with those who will be impacted by new technologies is essential in co-constructing relevant impacts and developing accountable algorithmic systems (Metcalf, *et al.*, 2021). Without the insight that could be provided by claimants, impacts were only constructed by the Labour Market Policy Group and the DWP's ethics committee. Furthermore, participants did not describe any mechanism for ensuring the DWP was accountable to the wider public regarding their consideration of potential harms which may develop from either the DTF or DPT projects.

In place of direct accountability to the public, participants expressed a feeling of accountability to the department, and the government itself, in addition to legal frameworks, as a proxy for citizens directly. While speaking about some of the difficulties on the project regarding data transfers, one participant stated:

> "So it's our responsibility to [citizens] to keep their data secured, and safely, that's why it's our responsibility as a department… but also, every analyst, or data scientist, we go through a rigorous process [to ensure the data is used for the right reasons]" (P1, Data Scientist)

Using Leonelli's distinction between responsibility and accountability, in which "responsibility [is] the moral obligation to ensure that a particular task is adequately performed" and "accountability denotes the duty to justify a given action to others and be answerable for the results of that action", participants' moral sense of responsibility was tied to the public, but their sense of accountability was tied to the organisation and the relevant regulatory frameworks (Leonelli, 2016). Thus, this suggests civil service practitioners and organisational structures may lack the necessary mechanisms for accountability towards those directly at risk of algorithmic harm.

Bias checking practices

Algorithmic bias mitigation frameworks suggest different ways of checking for bias, depending on the organisation's data, models, and organisational priorities (Jobin, Ienca and Vayena, 2019; Rovatsos, Mittelstadt and Koene, 2019). Primarily, guidance suggests conducting statistical checks which focus on either procedural (that is, ensuring that all data is treated in the same manner) or parity-based (that is, ensuring that all demographic groupings produce statistical outputs which are comparable equal) fairness assumptions. This section explores the type of bias checking which my participants were involved in within the context of the DTF project and Digital Plus Trial.

The wider goals of the DTF project were to provide a digital system which allowed for real time feedback for analysts running randomised control trials (RCTs). The Digital Plus Trial used a RCT method to determine whether some claimants would be suitable for a digital version of the traditionally face-to-face job-seeking service. Participants explained that RCTs are the standard departmental approach for testing policy interventions, and form the basis of many of the department's previous policy decisions. For example, previous RCT-based research on the effectiveness of conditionality informed the research design choices of the Digital Plus Trial.

When participants were asked about how they ensured the DTF project was fair, they referred to the perceived fairness of the RCT method. Further discussion revealed a strong faith in the method – which was seen as the 'gold standard' in the department – as long as it was set up properly. When asked how fairness and bias were addressed in the project's evaluation process, one of the Labour Market Policy Group analysts said:

> "I mean in terms of the [fairness and bias checking] approach we were taking, randomised controlled trial, so the randomisation process should have given us kind of groups who were pretty identical except the fact that one group of claimants would have the intervention, and the other group would have the business-as-usual service. So, on that side of things, we were kind of testing the randomisation process, just

> making sure that any kind of problems in terms of... the groups themselves with equally balanced" (P6, Analyst)

Using Green and Hu's (2019) conception of fairness-based approaches, this suggests participants were utilising a procedural understanding of fairness. By this, I mean an approach whereby it was perceived that by ensuring all individuals went through the same procedure, then the outputs from the Digital Plus Trial would be fair as a consequence of method. Additionally, participants also performed parity based statistical checks, meaning statistical checks which compared the outputs in different population groups, to ensure the treatment and control groups of the trial were receiving a similar level of sanctions.

When asked to recall any issues of bias on the DTF project, participants recalled the difficulties they had in augmenting the data collected from the UC database with the data collected from the participating Jobcentres using the exclusion questionnaire (also referred to as 'the tracker'). The data from the participating Jobcentres identified who was on the Digital Plus Trial, as the data science team was unable to do this using the UC data alone. Participants had observed that the treatment and control groups of Digital Plus Trial RCT had become unbalanced, so the baseline characteristics between the two groups were not comparable, meaning the results of the trial may have been influenced by these characteristics (such as age, gender, etc). One participant described the concern in the following way:

> "The hypothesis was whether it was control group work coaches not running the tracker when they needed to, 'cause they were just doing [the business as usual service], if they didn't run the tracker then who cares right, they didn't need the tracker for anything. [And] if I were a treatment work coach and [a claimant] came to me, and [they] said 'look you know I'm severely disabled and can't even speak English, then I might say 'well I don't need to run the tracker you aren't going to go on the trial' […] so it's plausible that work coaches in both treatment and control weren't running the tracker" (P5, Analyst)

Participants explained that this observation led them to investigate how data were being collected from the Jobcentre. Participants explained that during this process they found work coaches would forgo running the claimant through the exclusion questionnaire which the Jobcentres used to identify whether a claimant was on the Digital Plus Trial, if the work coach judged the claimant was not suitable for the trial (determined by an obvious disqualifying characteristic, for example if the claimant had an obvious disability, or struggled with English as a second language). This influenced how claimants were selected and assigned to the trial, leading to a selection bias. This illustrates the importance of data collection practices in bias mitigation, and additionally, how they are influenced by others in the organisation.

Participants had additional concerns about the fairness of the Digital Plus Trial. For example, one participant was concerned about the trial leaving more vulnerable claimants without the necessary face-to-face support they might require from a work coach, if they were in the treatment group of the trial and were not having success in their group search. He expressed concern these claimants may be "forgotten" (P4) and left without support from their work coach. This participant then said:

> "I think if you're not seeing that person in front of you, you might not have engaged with them with their because they're on the end of an email, you might not engage with them properly" (P4, Data scientist)

This suggests that the DWP may see it as a breach of ethics if claimants are not able to attend regular meetings with work coaches. While this participant was concerned about the fairness of a claimant missing out on support, it is worth noting that some research suggests many claimants find their work coach appointments negatively affects their mental health (Dwyer and Wright, 2014). This may suggest differences between how DWP practitioners view claimants' experiences of attending appointments with their work coach, and claimants' experiences themselves. This in turn suggests there may be challenges in practitioners judging the fairness of departmental interventions, a point which will be further discussed in Chapter 6 (paper three).

Additionally, participants discussed the reasoning behind the mechanics of the trial and who it was designed for. Cathy O'Neil argues that the construction of a mathematical model contains the designers' underlying assumptions (O'Neil, 2017). The inspiration for the Digital Plus Trial was the observation that some claimants required less assistance from a Work Coach than others, and "just need some money to tide them over" (PT4). According to my participants, in the very early development stages of the project, stereotypical descriptions of claimants were used in imagining potential claimant subgroups. The Digital Plus Trial was designed for those who the organisation saw as not needing support from a work coach, who can "just get on with it" (P3). The trial was designed with a very specific claimant in mind; one who is digitally capable, has a reasonable work history, knows how to apply for jobs, and requires little help getting back into work. The line of thinking was that if these claimants could take a 'digital journey', then work coaches could redirect their resources towards claimants with more complex needs.

The description of this claimant stereotype is consistent with the findings of Rosenthal & Peccei (2007), who describe how DWP work coaches construct different types of claimant subgroups. Moreover, they describe how work coaches construct the ideal claimant – one who is motivated to apply for work, needs very little help, and adheres to the department's ideological worldview (*ibid*). Additionally, Rosenthal and Peccei propose that this category of claimants is contrasted against 'bad' claimants; claimants who work coaches felt they were unable to help, either because their needs were too complex or because the claimant lacked motivation (*ibid*). Among my own participants, none suggested that any claimants were 'bad'. However, some of my participants categorised claimants as 'easy' and 'difficult' along very similar lines. When discussing potential for bias on the project, one data scientist said:

> "Not that I generally work with stereotypes, but I think stereotypes help, but the stereotype around the type of person whom the trial was designed to help and identify and support [were] people who do not need the continual support of an agent, they don't need a work coach, they don't need someone to tell them this is how you apply for a job, chances are they will get a job in 3-4 weeks, those were the people that we

> were trying to find, so [we could] spend our time on the people we know need a little more attention and a little bit more help" (P4, Data scientist)

While Cathy O'Neil (2017) focuses on how individual data scientists' opinions may become embedded within mathematics, this suggests the need to examine how organisational culture additionally embeds its biases within mathematical models.

The influence of organisational culture

This section explores the restrictions on model making which my participants described in their organisational context, specifically, how certain elements within their models were dictated by organisational practice.

Several academics have argued that algorithmic bias may develop from the strategic goals of a project by embedding and re-enacting ideological discourses as part of the design process (Eubanks, 2018; Costanza-Chock, 2020; Green, 2021). Within the department, participants described very clear aims the department worked towards. One participant commented:

> "When you're working with DWP it's always the goal to get people closer to the labour market, into a job and self sustaining... " (P2, Data scientist)

The department's overarching philosophy very clearly prioritised 'getting claimants back into work', an aim which visibly guided my participants' objectives and working practices. Once again, this echoes Rosenthal & Peccei (2007)'s findings from interviewing DWP work coaches, in which they found participants adhered to an overarching philosophy which strongly valued getting claimants into work above other potential outcomes. Indeed, my participants explained how this goal provided both the impetus for the Digital Plus trial – by investigating whether work coaches' time could be 'freed up' to spend with claimants who found it harder to find work – in addition to being the basis of The Treasury's major concerns surrounding the trial. During the trial, there was concern from The Treasury that the proposed trial, and any policy which came from it, may negatively impact claimants becoming closer to the job market, due to its potential impact on the level of conditionality designed into the UC system. It was implied that The Treasury were concerned that participants in a digital version of the 'business as usual' service would not meet the same level of conditions to receive their welfare payment as those on the face-to-face version of the service. Participants explained that the Treasury's reasoning was based on previous research conducted by the department which had found that "[welfare] conditionality works" (P1) in getting claimants back into work. For this reason, much of the time designing the trial was spent on ensuring that conditionality was preserved in the digital arm of the trial. Or, to put it another way, that claimants on the digital group of the trial (the treatment group) were still keeping up with their 35 hours' worth of job-seeking activities a week and could be sanctioned accordingly if they did not. One participant explained:

> "[The] Treasury held us to account because we had strong RTC evidence that conditionality in general worked, I mean it's also a consideration that we want to make, in terms of if we thought it wasn't working, we could switch the trial off and do it quickly. " (P5, Policy Analyst)

My research participants were constrained, in terms of what they did on the DTF project and Digital Plus Trial, by the rules already in place at DWP, including Universal Credit conditionality, in addition to the department's governing philosophy of getting people into work. Indeed, one of my participants explained that this was one of the benefits of the DTF system being developed for digital trials – they would be able to analyse these trials in near real time and shut them down if they negatively impacted either the claimant or the department's goals. As previously mentioned, the data science team performed comparative statistical analysis to assess whether the treatment and control groups of the trial were receiving a comparable level of sanctions. Due to The Treasury's concerns around the loosening of conditionality, the team needed to prove that claimants in the digital group of the trial were being sanctioned comparably to those in the face-to-face group, to ensure the groups were being treated equally. Furthermore, participants described this as being part of the condition for the DTF project / Digital Plus Trial to be allowed in the first place:

> "We had to get permission from Treasury to do all of this stuff, because we get funding from Treasury and our previous labour market evidence on conditionality is all in a face to face setting" (P5, Analyst)

Rodger (2012) argues that welfare systems which enforce conditionality allow the state to exert its power to compel claimants into behaving in a specific way in order to qualify for assistance with materially necessary goods such as food and housing (Rodger, 2012). Furthermore, he asserts that conditionality is ideologically aligned with systems that support discourses framing those in receipt of benefits as being either the 'deserving' or 'undeserving' poor. A more complex picture of the efficacy of conditionality is provided by Dwyer *et al.* (2020) on the experience of claimants. They found that claimants with mental health difficulties find conditionality either largely ineffective at helping them into work, or that it negatively impacts their work prospects (Dwyer *et al.*, 2020). While my participants mentioned that RCTs were considered the 'gold standard' within the department, due to their perceived objectivity and reliability, little attention was paid to political or ideological biases within the trial process. Instead, participants expressed the opinion that the preference for RCTs was due to the objectivity of the method. However, Monaghan and Ingold suggest evidence based policy outcomes are never purely evidence based – they are in part influenced by political will and feasibility (Monaghan and Ingold, 2019). Additionally, while RCTs are a popular choice for public sector data analysis, they have been criticised for their use in policy situations where there is a lack of clarity around the epistemological assumptions on which RCTs rely (Deaton and Cartwright, 2018). Specifically, RCTs are particularly vulnerable to a lack of external validity (whether results can be generalised outside of the trial environment).

This suggests that in addition to the issues presented about using procedural fairness methods, mentioned in section 4.2, further complications may arise from pre-existing evidence relied on by the department.

4.5. Discussion and conclusion

This paper investigated the research questions; **What algorithmic bias working practices are currently practiced by data science practitioners? What are the limitations of these practices?** This was explored by interviewing data science practitioners and policy analysts at the DWP, and focusing on an active data science project developed to allow practitioners to perform real time RCTs. This study contributes towards understanding algorithmic bias mitigation guidance in two ways. Firstly, my participants relied strongly on legal frameworks to provide bias mitigation guidance, due to their position as civil servants. Participants perceived adhering to legal frameworks and government processes as fulfilling their duty to the public they served. Furthermore, while participants' moral sense of responsibility was tied to the public, their sense of accountability was tied to the organisation and the relevant regulatory frameworks. Secondly, my participants described how pre-existing organisational knowledge production limited what approaches might be considered to mitigate algorithmic bias. Specifically, participants described how the team was expected to incorporate the findings of previous research regarding the success of conditionality into the Digital Plus Trial project. I expand on these below.

Regarding the first finding, in common with the insights provided by Orr and Davis (2020), the practitioners I interviewed strongly relied on legal frameworks within the context of their work (Orr and Davis, 2020). This was facilitated by organisational practices such as legal training, and standardised processes to ensure the legal requirements associated with departmental data science projects were conducted appropriately. These processes paid careful attention to GDPR regulations. However, while Orr and Davis (2020) perceived the private sector practitioners in their sample as adhering to legal frameworks to offload the moral responsibility within the context of their work, my research found a different experience for civil servants. Using Leonelli's distinction between responsibility and accountability, in which "responsibility [is] the moral obligation to ensure that a particular task is adequately performed" and "accountability denotes the duty to justify a given action to others and be answerable for the results of that action" (Leonelli, 2016, p3), participants' moral sense of responsibility was tied to the public, but their sense of accountability was tied to the organisation and the relevant regulatory frameworks. Instead of using legal frameworks to offload [their moral responsibility, practitioners performed their duty to citizens through adherence to legal and political structures, which sometimes required divestment of their own personal stance on moral decision making.

Regarding the second finding, I argue that practitioners' working practices in relation to bias checking were limited by previous research conducted by the DWP, in addition to influences from the department's organisational culture. My participants described how the teams'

project goals, modelling methods, and assumptions were stipulated by the Treasury to fit into the department's previous evidence base. This was particularly notable regarding the influence of the department's existing research on welfare conditionality. However, the effectiveness of this mechanism has been contested by researchers outside of the DWP, who suggest that conditionality may negatively impact some individuals' ability to find employment, in addition to increasing the likelihood of housing insecurity (Dwyer and Wright, 2014; Williams, 2022). Moreover, to ensure conditionality was preserved on the project, sanctions were used as a proxy to measure the level of conditionality present in both groups of the trial. For this reason, the trial presupposed that claimants experiencing a digital version of the UC service would not improve adherence to their work search commitments.

Additionally, despite my participants' strong concerns about methodological rigour, they expressed an awareness that the foundations of the trial were somewhat based on stereotypes, rather than on empirical research about claimant subgroups. These conceptualisations parallel those found in research by Rosenthal & Peccei's (2007), who identified that work coaches judged claimants as either fitting the 'good' or 'bad' claimant stereotype. This parallels Eubanks' (2018) description of how discourses of the 'deserving' and 'undeserving' poor perpetuate decisions to restrict the resources given to, and justify different treatment of, poor people who are judged to not be worthy of societal aid. Eubanks argues that categorisation of the poor into these two categorises allows governments and services to moralise their lack of support, or their increased surveillance, of citizens who are judged to be 'undeserving'. Like Eubanks, I suggest the construction of claimant types for the purpose of statistical modelling may perpetuate prejudice against particular claimants. It is necessary for practitioners to carefully consider how these stereotypes may enforce prejudice against particular claimants within an organisational context.

This study contributes towards understanding the areas discussed in algorithmic bias mitigation guidance in two ways. Firstly, my participants strongly relied on legal frameworks due to their position as civil servants, but the legal frameworks to which they were required to adhere did not facilitate accountability to the population they served. Rather, while my participants felt a responsibility to the public, they were accountable to the state. Secondly, my participants' working practices in relation to bias checking were limited by previous research on conditionality conducted by DWP, in addition to influences from the department's organisational culture. This was present in two prominent ways; first, in how the Treasury stipulated the project must be run – it had to preserve conditionality, due to its positive assessment in the department's previous research in this area. Second, the trial was designed using stereotypes of what constitutes a 'good' claimant.

Additionally, this paper contributes towards answering my thesis' overarching research aim: to investigate how might DWP practitioners mitigate the impacts of algorithmic bias. The findings presented in this study suggest the civil service context has its own unique challenges in algorithmic bias mitigation, and further research is required to assess what methods may be appropriate for this context. Additionally, it suggests civil service practitioners already strongly rely on the rules and guidance regarding organisational practice, which presents an

opportunity for further algorithmic bias mitigation efforts. These will be explored further in Chapter 5 and 6.

Chapter 5: Paper 2, Lessons in mitigating bias from the field: exploring good practice and moral challenges on the AuroraAI project

In the previous chapter, I presented the first of my three thesis papers. This paper addressed the research questions: what algorithmic bias working practices are currently practiced by data science practitioners? What are the limitations of these practices? These research questions were designed to discover DWP's current algorithmic bias mitigation working practices, as a basis for the subsequent research papers. I found that my participants strongly relied on legal frameworks to provide bias mitigation guidance, due to their position as civil servants. Participants perceived adhering to legal frameworks and government processes as fulfilling their duty to the public they served. Secondly, my participants described how pre-existing organisational knowledge production limited what approaches might be considered to mitigate algorithmic bias. Specifically, the department required data scientists to incorporate previous findings from their in-house evidence-based research about conditionality into the models they produced, which may have led to the adoption of certain assumptions around those claiming social security benefits.

The findings from Chapter 4's study suggest the civil service context has its own unique challenges in algorithmic bias mitigation, and that further research is required to assess what methods may be appropriate for this context. This finding led to the development of the second research paper, which is presented in this chapter. This paper focuses on what civil service organisations which might be considered to be more advanced in their thinking on algorithmic bias were doing to mitigate these risks. The insights identified in this paper would form the basis of the final research paper (Chapter 6), which would investigate how these insights might be adopted in a DWP context. In this chapter, the second paper is presented.

5.1. Introduction to paper two

As already noted, in recent years, algorithmic technologies have received widespread criticism of their discrimination against marginalised groups and the way in which they further entrench historical inequalities. Increasingly, these types of cases have been seen in the public sector, for example, in how assessment scores used in the criminal justice system falsely report black defendants as having a higher chance of reoffending (Angwin *et al.*, 2016); how child abuse assessments accuse people of child abuse for existing in the material conditions of poverty (Eubanks, 2018); and how less privileged teenagers received lower algorithmically calculated grades on the basis of them attending a less well-performing school (Smith, 2020). This phenomenon, known as 'algorithmic bias,' describes how through a combination of social, technical, and probabilistic mechanisms, some people are penalised, or denied opportunities, due to their membership of a marginalised group. While algorithms are adopted by public sector organisations in an attempt to enhance and improve public

services, instead they often bring further difficulties, in addition to questions about whether these technologies should be implemented at all.

There has been considerable debate around how 'good practice' might be defined when attempting to mitigate algorithmic bias. As academics and practitioners have looked towards mitigating the risks of algorithmic bias, a wide selection of framings has been adopted in an attempt to further this goal. These have included ethics, fairness, non-discrimination, and justice, as discussed in Chapter 2. Within the context of this paper, I primarily use the term 'ethics' to describe structured organisational approaches of ensuring social good outcomes of technological developments and processes.

While much of the literature around algorithmic bias thus far has discussed the harms caused by algorithmic technologies (Redden and Brand, 2017; Eubanks, 2018; Dencik *et al.*, 2022), less is known about what practical steps public sector services are taking to mitigate risks of algorithmic bias in the design and implementation of algorithmic systems. To address this gap, I investigated the following research questions: What might 'good practice' on an algorithmic project look like? What challenges does good practice on an 'ethical AI' project face in practice? Understanding how practitioners engage with and implement good practice, and the successes and difficulties which arise therein, is important in order to illuminate how designers can grapple with the difficulties posed when working within complex socio-technical environments. To investigate this issue, I collected qualitative data through semi-structured interviews and utilised document analysis to better understand how stakeholders on the AuroraAI project by the Finnish Ministry of Finance are responding to the challenges posed by algorithmic bias. Additionally, I interviewed AI Ethics experts, predominantly from algorithmic justice organisations, about the AuroraAI team's proposed algorithmic bias mitigation plans. The AuroraAI project was chosen as a unit of analysis due to its reputation as a progressive AI project and the public availability of the project's ethical principles. The term 'AI ethics experts' here describes actors who work towards developing structured approaches to ensuring social good outcomes of technological developments and processes.

In this study, I identify two key findings regarding algorithmic bias mitigation. First, the findings suggest that even in this purportedly progressive project, there is a lot of disagreement about what constitutes good practice in mitigating algorithmic bias, and the types of solutions that might be practically implementable. Participants operationalized different definitions of key algorithmic bias terminology, which brought with them tensions around how to communicate about the project's risks and resolve ideological conflicts. Second, while all participants on the AuroraAI project were committed to developing a public sector AI system which improved society, challenges arose as a result of fast-paced project management styles. In this environment, project goals which reinforced cultural assumptions regarding individual responsibility were left unexamined.

The rest of the paper is laid out as follows: it begins with a review of the literature in this area, and details recent key developments in public sector algorithmic technologies and algorithmic bias mitigation, drawing on literature that focuses on these as design challenges. I then

provide a breakdown of the key terminology used to discuss 'good practice' in relation to algorithmic bias mitigation, including ethics, fairness, and justice. I subsequently detail my methodological approach to selecting a project for study, my data collection process, and explain how these are suited to investigating my research problem. Lastly, this is followed by the findings of the study, as well as a discussion of how these contribute towards current academic understanding of addressing algorithmic bias in the public sector.

5.2. Literature review

Designing algorithms in the public sector

As in many industries, public sector organisations have recently started adopting algorithmic technologies. In some cases, this has meant organisations relying on third-party services, and in others, designing and developing their own algorithmic systems in house. In the context of the UK for example, Durham Police Constabulary developed the HART (Harm Assessment Risk Tool), which was designed to assess the likelihood of criminals reoffending, using machine learning techniques based on the organisations' own historic arrest data (Oswald *et al.*, 2018). In contrast, instead of relying primarily on in-house data, public sector organisations such as Kent County Council, The London Fire Brigade, and Lancashire County Council have used either Experian's Mosaic database, analytical tools, or both, to provide them with workflows which allow them to profile and target support to different groups within their care or remit (Dencik *et al.*, 2018). Both methods have been criticised as perpetuating biases or stereotypes about people based on location or demographic data (Big Brother Watch, 2021) (for more information about how this occurs, see Chapter 2). In addition to discrimination concerns, systems strongly relying on third-party data have been criticised for upending the safeguards of democratic institutions, because these third parties are not subject to the democratic oversight embedded in public sector organisations (Balayn and Gürses, 2021). Furthermore, there are concerns that this development economically entangles the public sector with algorithmic tool providers and Big Tech companies (Balayn and Gürses, 2021). Therefore, enmeshing public sector practice with powerful organisations with the motivation to provide returns to their shareholders, rather than provide services for the public themselves (*ibid*).

As discussed earlier in this thesis, many of the proposed solutions to algorithmic bias have hitherto been technical or procedural in nature. These have included debiasing techniques which focus on making datasets either more representative of their target population or which rely on achieving statistical parity when comparing the outcomes of different groups based on protected characteristics (Galhotra *et al.*, 2017). Other approaches have included finding new ways to operationalize the concept of fairness within a statistical framework, allowing practitioners to better perform statistical checks on their models (Bellamy *et al.*, 2019) (As discussed in Chapter 2: Technical mitigation methods). However, these methods have been criticized by some third sector organizations and academics for not addressing the social structures and processes that contribute to algorithmic bias (Hoffmann, 2019; Balayn and Gürses, 2021).

Concern about discriminatory biases in computing systems is not new. Indeed, Friedman and Nissenbuam's (1996) early paper on bias within computer systems identified bias as a key design concern. Moreover, they called for 'freedom from bias' to be considered a design ideal within these systems – in the same way that reliability, accuracy and efficiency of computer systems are within the computing community (Friedman and Nissenbaum, 1996, p346). Recently, scholars have furthered this concept, notably Costanza-Chock, whose Design Justice framework discusses how technology can "reproduce and/or challenge the matrix of domination (white supremacy, heteropatriarchy, capitalism, ableism, settler colonialism, and other forms of structural inequality)" through the developers' design choices (Costanza-Chock, 2020, p23; Hill, 2000). Their work takes on an intersectional lens to understand the complexities involved when designing inclusive technologies. Furthermore, these approaches seek to judge technologies by the outcomes experienced by those impacted by these systems, not the often-well-meaning intentions of the designers in question (*ibid*).

Park and Humphry (2019) explored how the participatory approach of co-design was used to create the Australian Nadia chatbot, which aimed at assisting disabled citizens with queries, providing them with a service which aids them in overcoming accessibility barriers when interacting with public sector services. The chatbot was created using co-design principles from the outset – specifically to avoid creating a service which further discriminated against their already marginalised users. This aspect of the project was well received by citizens who participated, and had good levels of engagement from the target community. However, this was difficult in view of the tensions that exist in a public sector context, between the need for a highly predictable service, and the inability to feed the chatbot enough data to ensure this predictability prior to launching the chatbot, and consequently the project never got off the ground (*ibid*). This project also highlights the sociotechnical complexities public sector algorithm designers engage with when attempting to implement good practice principles.

Design philosophies such as participatory design suggest mechanisms to minimise algorithmic bias (Costanza-Chock, 2020). Participatory design approaches focus on creating design processes where the intended users or stakeholders take a critical role in designing the project (Schuler and Namioka, 1993). Similar frameworks have been suggested by disability justice activists through slogans such as "nothing about us, without us" (Costanza-Chock, 2020). Practically speaking, participatory approaches suggest instruments which allow users to engage in the design process, either directly, or via, user groups, surveys, hackathons, or other mechanisms designed to foster accountable relationships between designers and those affected by the decisions made by these systems.

However, approaches such as these have been subject to criticism. A prominent concern regarding these approaches is that often designers do not involve end users in a way which is truly meaningful, and which provides users real opportunity to influence the goals of a project; rather, often they merely provide a forum for users to provide feedback which may or may not be acted upon (Sloane *et al.*, 2022). In recent years, this has been called 'participation washing' (Sloane, 2020). Other challenges can include the difficulty in creating the necessary relationships for this type of work, especially if the community for which the

technology is being designed was reason to mistrust the designers (Neville and Weinthal, 2016).

Designers utilise co-design approaches as it is believed that these methods can help them achieve the social good. However, there are questions regarding how designers may conventionalise 'the social good.' In the following section, I examine how designers define 'good practice' and 'the social good.'

Defining 'good practice' and the 'social good'

Alongside the difficulties designers have when attempting to actualise good practice frameworks, there has been considerable debate around how 'good practice' might be defined when attempting to mitigate algorithmic bias. In this section, I explore the different ways the 'social good' has been conceptualised in this space. This is discussed through an examination of the key terms often used within both academic and public discourse in relation to algorithmic bias mitigation. The key terms discussed here are 'ethics', 'fairness' and 'justice'. The definitions of these can be circular when used in lay speech, and they also go through fashionable turns academically (Sayer, 2011, p16). However, they have distinct connotations within current algorithmic bias discourse. I briefly discuss the ways in which these words have been defined within the discussion around algorithmic bias mitigation, and additionally qualify how I will be using them in this paper.

Ethics

Ethics is generally understood as being the study of moral principles and how these should guide personal and societal behaviour. Academically, the discipline of philosophy has provided much of the historic groundwork for the different schools of thought in this area (refer to Chapter 2). However, in recent years, there has been a surge of organisations adopting what is known as 'organisational ethics' or 'business ethics' approaches to dealing with the moral dilemmas found in organisational working practices (Vogel, 2006; Mckinsey & Company, 2022). This has been influenced by the rise of 'corporate social responsibility,' which began in the 1990s, through which organisations began focusing more on what types of ethical practice were practical and affordable, in addition to practicing greater 'stakeholder engagement' (Vogel, 2006).

When organisational ethics are applied in the context of designing algorithmic technologies, it has been defined as encompassing a) the moral consideration with which designers build these technologies, and b) the moral decisions these technologies will be programmed to make (Wing, 2018). Moss and Metcalf's (2021) report on Silicon Valley ethics workers (described as "ethics owners") describes a shift from the older form of ethics work done in these companies, in which ethics workers deflected public pressure and demonstrated legal compliance, to the ethics roles of today which focus more strongly on preventing social harm (Metcalf, *et al.*, 2021). This type of ethics work involves navigating stakeholder relationships and market pressures, and managing company resources (*ibid*). Additionally, these roles focus on preventing harm within the parameters set by the business, with issues occurring when

ethics workers try to work beyond these limits. For example, when Timnit Gebru in her capacity as an ethics lead at Google co-authored a paper detailing the risks that utilising large models will exacerbate carbon emissions, have unknowable biases, and spread misinformation, she was subsequently fired by the company (Hao, 2020; Bender *et al.*, 2021).

One of the challenges of the business ethics framework is the lack of agreement between ethics workers as to what constitutes 'ethics', with different practitioners bringing with them different sets of personal moral codes to their work, making it difficult for ethics workers to create organisational processes which are workable and consistent (Moss and Metcalf, 2020). Moreover, this may lead ethics workers to focus their attention on more quantifiable benchmarks such as bias or fairness measures (*ibid*). Within this paper, I chiefly use the word 'ethics' to describe the processes organisations use to work towards some form of 'social good' or to otherwise address moral dilemmas within the context of their organisation, such as through ethics boards, procedures and policies.

From fairness to justice

While in layspeech the word 'fair' is often used to mean something which is free from bias, dishonesty, or injustice (Dictionary.com, no date), it has taken on specific connotations within debates on algorithmic bias. When defined and operationalised by data practitioners, who are often core designers of the systems being discussed, two prominent operational definitions of fairness have emerged; procedural fairness and statistical fairness (Green, 2018; Green and Hu, 2018). These definitions have been discussed in detail in earlier chapters in this thesis. As noted above, while both definitions capture important aspects of fairness, they primarily focus on fairness as a result of method. What they fail to consider is the fairness of the outcomes when regarding the wider socio-economic system within which the data and its outcomes are embedded (*ibid*). Furthermore, they do not consider how these mechanisms interplay and intersect with other social structures to reproduce unequal outcomes.

Hoffman (2019) argues that striving for fairer algorithmic systems does not go far enough, and a goal of producing 'fairer' systems leads designers to focus on 'bad' algorithms and 'bad' data. She argues that this creates a narrow field of inquiry, which limits practitioners' ability to recognise how data and algorithms connect to wider issues of injustice within the context of society (Hoffman, 2019; Costanza-Chock, 2018). For example, if designers only focus on ensuring an algorithm used in the criminal justice system judges a defendant's likelihood of reoffending in a way which is 'fair,' by ensuring there are an equal number of high risk offenders regardless of race, this process would not address the wider injustice of incarceration. Not only does this narrow the practitioners' focus away from how discrimination is structural in nature, but it also externalises the biases leading to discrimination, minimising the connection between our internal experiences and assumptions and the society and structures around us (*ibid*). Furthermore, it de-emphasises the non-static nature of social structures, furthering the idea that discrimination is one thing in particular, rather than highly contextual and subject to change. Hoffman advocates approaches which consider these contextual issues, describing these as 'justice' based approaches (*ibid*).

As discussed in Chapter 2, in *Towards Data Justice*, Dencik *et al.* (2016) argue that the issue of algorithmic bias should be understood through the framework of justice, not fairness (Dencik, Hintz and Cable, 2016). They suggest that a justice-based framework connects the issues regarding data-driven practices to those of inequality and exploitation more generally. Additionally, they suggest that a justice-based framework provides a conceptual foundation for creating tools to address these issues. It does this, they argue, by providing a way of examining the ideological basis of data-driven processes and considering power relations, interests, and political agendas within the context of data-driven practices (*ibid*). From there, it provides a foundation to question how society should be organised (Dencik *et al.*, 2022; Milan & Treré 2019, 2021; Treré 2019).

5.3. Methods

The aim of this study was to investigate what 'good practice' on an algorithmic project might look like, the challenges which good practice on an 'ethical AI' project face in practice, and how these are overcome. To explore this, I investigated an algorithmic project in development by a civil service organisation, the AuroraAI project by the Finnish Ministry of Finance, to understand their attempts at implementing algorithmic bias mitigation approaches.

I collected qualitative data through semi-structured interviews with practitioners and experts actively working in different disciplinary capacities on AuroraAI. Furthermore, I utilised document analysis to understand how stakeholders on the AuroraAI project were responding to the challenges posed by algorithmic bias. Additionally, I conducted semi-structured interviews with AI Ethics experts outside of the AuroraAI project, predominantly from algorithmic justice organisations in the UK, the European Union, and USA, about the AuroraAI team's proposed algorithmic bias mitigation plans and to understand the wider context of algorithmic bias mitigation efforts in the public sector. The AuroraAI project was chosen as a unit of analysis due to its reputation as a progressive AI project and the public availability of the project's ethical principles.

The research was conducted in four key stages: 1) I conducted desk research on public sector algorithmic projects against a set of criteria to create a shortlist of potential projects on which to focus my study. 2) After identifying a suitable project, I used document analysis techniques on the project's public facing documents to understand the algorithmic project's algorithmic bias mitigation methods, ethical commitments, goals, and project structure. 3) I conducted semi-structured interviews with people working on the project. 4) I conducted semi-structured interviews with AI Ethics experts discuss the AuroraAI team's proposed algorithmic bias mitigation plans and to understand the wider context of algorithmic bias mitigation efforts in the public sector. The experiences of these experts are well placed to offer insight into both the wider structural problems and the organisational challenges faced, when attempting to implement good practice in relation to algorithmic bias mitigation. I expand on these stages below.

As noted in Chapter 3, to select a suitable project, I conducted desk research to find out what algorithmic public sector projects were being developed at the time of the research. The objective was to find a public sector project taking strong steps to mitigate the risks of algorithmic bias. The criteria for this were as follows. First, the project must have publicly posted their ethics guidelines, which had to include consideration of algorithmic bias. Furthermore, the project had to be at a stage where some technical development had started to take place. Other criteria included the need for the project to process data relating to a person, as prior research on the issues of algorithmic bias is restricted to this type of data processing. On top of these considerations, there were also accessibility considerations. As the only researcher on the project, I would only be able to interview practitioners in English. To look for a project which fit these criteria, I used a combination of internet search methods in conjunction with the Oxford AI readiness index, to pick countries which were more advanced in this area (Oxford Insights, 2019). Additionally, I searched public algorithmic registers such as the Helsinki algorithm register. These criteria proved very restrictive and left the most suitable project as the AuroraAI project by the Finnish Ministry of Finance.

5.4. Description of AuroraAI project

The AuroraAI project is designed to provide citizens with recommendations as to what services they might benefit from depending on what 'life event' they are currently experiencing. Life events include situations such as getting a divorce, moving to a new city, or starting higher education. The 'life event' approach is an established eGovernance framework within the Finnish public sector (Gros, 2020). This approach aims to allow government services to easily catalogue citizen queries for both government and citizen use, with 'life events' acting as "structuring metaphors" (*ibid*).

AuroraAI's recommender system planned to utilise citizens' demographic and lifestyle data to provide personalised results when using the service. Much like virtual assistant applications such as Apple's Siri, AuroraAI was expected to be able to provide 'intelligent' recommendations as to what services would best suit the individual when asked specific queries. The interface for submitting queries relied on chatbots, which use natural language processing to parse meaning and generate responses. In addition to using the 'life events' framework, AuroraAI relied on the concept of 'digital twins' which a concept designed to allow users to control the data they input into the application as well as their preferences with respect to when information should or should not be shared with specific organisations (Jones *et al.*, 2020). The AuroraAI system then combined this 'life event' data with the data shared by the citizen in their 'digital twin' and used these to create personalised recommendations for their circumstances and characteristics. In addition to the recommender system, there were plans to expand AuroraAI to include a 'fitbit' style mechanism. This would act as an assistant and dashboard which encourages the user to work towards their own customisable goals, such as changing career or improving their health.

From a project management standpoint, the AuroraAI project consisted of a network of smaller pilot projects, each project managed separately by its overseeing organisation. The overseeing organisation might be a private company, a public sector service, or any organisation which would be interested in producing a pilot based on a specific 'life event.' Since the project's conception in 2018, the Finnish Ministry of Finance arranged for several small pilots to be produced based on different life events, and these were used to test and develop different aspects of the AuroraAI system. At the time of this research, the pilot projects operated independently of each other.

During the second stage of my research, I analysed documents to understand the AuroraAI team's approach to mitigating algorithmic bias, as well as the context in which the project was being developed. Document analysis is particularly well suited to studies which focus on a well-defined context as the unit of analysis, due to its ability to provide understanding and rich details, and to uncover new meanings throughout the process (Bowen, 2009). To do this, I analysed the project's public facing documents available in English, which included the AuroraAI roadmap, the 2020 AuroraAI launch conference, the Sumodigi Podcast series discussing AuroraAI, and the AuroraAI development documents. The analysis of these documents informed the creation of the interview guide used in the third stage of the project.

In the third stage of the study, I conducted six semi-structured interviews with practitioners working on the AuroraAI project. These participants had roles ranging from special advisors, data engineering, data work, ethics expert, research co-ordinator, pilot project leader, project tech lead, and prison worker. Some participants on the AuroraAI project held more than one role. Participants were approached in various different ways, including emailing key figures mentioned in the AuroraAI documentation, posting an invitation on the AuroraAI slack channel, and snowball sampling (Bryman, 2004). Interview questions included topics such as asking what role the participants had on the project; why they were interested in it; how the project fitted into the broader environment of service digitalisation initiatives within Finland; how they intended to measure the success of the project; what they considered to be good practice ethically; how they intended to mitigate algorithmic bias on the project; any challenges in mitigating algorithmic bias; and what steps would be taken to ensure this throughout the project.

In the fourth stage of the study, seven practitioners working in AI ethics, either at algorithmic justice focused organisations or otherwise engaged in AI ethics work, were interviewed using semi-structured interviews. Participants were asked questions relating to the AuroraAI teams' proposed algorithmic bias mitigation plans, and their experience of public sector projects which had been successful in mitigating algorithmic bias. These participants were included within my sample for two reasons. First, after interviewing the AuroraAI participants, it became clear that the project was not as far along in implementing their ethical principles as their ethics documentation had suggested. This made it prudent to find insight into other organisations' ideas about good algorithmic bias mitigation. Second, by including AI Ethics experts outside of the AuroraAI project, I would be able to identify differences in approaches in the wider algorithmic bias mitigation environment.

Interviews were held using video platforms between July and September 2021, with some interviews having a follow up discussion via email. Interviews lasted between 40-120mins. Prior to starting the data collection process, ethics approval was sought from the University of Sheffield, which considered issues around participants' privacy, confidentiality, and data management. Both the interview transcripts and project documents were analysed using reflexive thematic analysis (Braun and Clarke, 2006). For further information on my analysis process, see Chapter 3 (Methodology) section 3.8.

5.5. Findings

The findings below focus on four themes. The first, 'differences in imagining good practice', focuses on how practitioners differently conceptualised what good practice might be in an algorithmic project. The second, 'operationalizing good practice', focuses on how practitioners sought to implement good practice concepts. The third, 'project communication and project management', focuses on how project communication and management styles influenced algorithmic bias mitigation efforts. And lastly, 'difficulties envisioning change' focuses on how the wider social context within which projects were developed impacted practitioners' ability to envision change. I expand on these below.

Differences in imagining good practice

Prior to implementing good practice, practitioners require some conception of what good practice is. This section explores how practitioners differed in how they imagined good practice. These include differences between how those on the AuroraAI project and outside AI ethics experts imagined good practice, and how practitioners with different roles on the AuroraAI project viewed good practice. The differences presented here include at what part of the project's life cycle ethics work should begin, the expectations around what an ethics committee should do, and the differences in value-based assumptions held by practitioners.

Participants had different ways of conceptualising good practice in relation to ethics work. The primary instrument for mitigating the risks of algorithmic bias on the AuroraAI project was through an ethics committee which provided feedback on the project's development. However, participants disagreed about when ethics work should commence during the project's life cycle. One member of the ethics board, a senior advisor on IT governance, lamented the lack of any ethics work conducted at the very start of the project, and recounted how he had emphasised to the AuroraAI team the need for an ethics committee to be set up for the project. He explained how this process had involved rounds of attempts to convince them of this. He noted that, as the project's senior stakeholders originally saw very little risk associated with this type of project, they therefore did not believe it required an ethics committee. Eventually, they were convinced, and agreed to provide the resources to set up the project's ethics committee. However, by the time this occurred, the committee felt uncertain as to what influence they could have when all the project's major goals and principles had been set:

> "What influence can we even have at this point when all the major goals have been set. The major technologies have been decided on. Why are we here? […] I think we're there to try and do what we can at this point, it remains to be seen what we can actually do" (P5, AuroraAI Ethics Committee).

Additionally, there were tensions on the AuroraAI project between the ethics committee and non-ethics committee members, with both groups having different expectations of the ethics committee's role. One participant, whose role involved data work, said he would like the ethics team to come in and 'diagnose' ethical issues they were struggling with, such as why an algorithm was biased, and to offer practical solutions. However, he had so far found the ethics committees feedback to be difficult to translate into action or the types of processes he was usually presented with. The participant from the AuroraAI ethics committee, and participants from outside AI ethics organisations, felt that not all problems fitted within a technological lens. Indeed, most of these participants would often emphasise the need to analyse issues of algorithmic bias by considering how the aim of any proposed technology relates to, and may exacerbate issues, as part of the wider social structures within which the technology is embedded. However, on the AuroraAI project, the ethics committee's non-technological conceptualization of bias seemed to further a gap in understanding between the ethics committee and data practitioners.

Comments from some of the AuroraAI data practitioners seemed to fall into the 'framing trap'. Selbst *et al.* (2018) describe an abstraction trap as an error made when computer scientists "[abstract] away" aspects of the social context as part of the process of constructing a model. One of the abstraction traps described by Selbst *et al.* (2018) is the 'framing trap', which describes how practitioners attempt to solve social issues such as 'fairness' using the methods already at hand, for example by making different modelling choices or using de-biasing techniques with the currently available data. Data practitioner participants on the AuroraAI project sometimes found it difficult to see beyond an algorithmic or datafied framing of the problem at hand (Selbst *et al.*, 2018), and seemed to expect the 'answers' given to them to fit into an algorithmic or datafied framing. This seemed to be due to the wider dynamics at play during the project, whereby practitioners felt pressure to produce results quickly (described more fully in Findings: Project Communication and Project Management).

Furthermore, those in senior management positions on the project conceived ethics work and working towards the social good differently to both AuroraAI data practitioner participants and the participant from the ethics committee. When reflecting on the purpose of the ethics committee, one of these senior management participants commented:

> "That kind of ethical group need to be constructive, so basically, they shouldn't be like a rock in the shoe. Basically, it should be more like constructive builders, so they try to give some sort of an idea, so, that is the direction, if we are going to that direction, then there might be these kinds of challenges and these kinds of challenges." (P3, AuroraAI senior management)

While he stated that he wanted the project to be based on strong ethical foundations, the comment that these "shouldn't be like a rock in the shoe" implied there was concern that an

ethics committee can hinder a project's momentum and efforts towards producing tangible results. This echoes comments from Moss and Metcalf's (2020) findings in a Silicon Valley context; that some practitioners are reluctant to engage with ethics workers because of the perception that it prevents them from getting on with their work . Moreover, this suggests the different expectations of people in different roles influence how they engage with ethics work in a professional context, with senior stakeholders potentially more concerned about the prospect that an ethics committee could hinder a project delivering tangible deliverables.

In addition, senior stakeholder participants were more concerned about issues regarding data security, privacy, and organisational transparency than algorithmic bias. It seemed that more of the project's technological resources were put towards those issues, rather than algorithmic bias, due to the long shadow cast by large tech corporations' reputation for carrying out unnecessary surveillance. One senior member of the team said, in relation to his perspective on the ethical expectations of the project:

> "Maybe my biggest concern is that if we are not going to figure out how we should do this age of AI in an ethical way, then we have great challenges ahead. And as you know, the greatest, greatest challenges of our time comes from maybe the private entities, the private organisations who are basically gathering all the data around their own assets." (P1, AuroraAI tech lead)

These concerns about large private organisations' data collection practices influenced the choice to use digital twins on the project, with the aim of allowing citizens to control the data about them which is shared with the AuroraAI system. Approaches such as these have received a mixed reception – some claiming these empower citizens to make their own choices regarding how their data is used, and others claiming these approaches place an unfair burden on individual citizens. Regarding the first claim, Hartmen *et al.*'s (2020) study into public perception of good data management found that UK survey respondents liked the idea of having more control over their data, with approaches similar to digital twins, and preventing private organisations from profiting from the use of their data. However, regarding the second claim, the authors also stress that this area is complicated, as previous qualitative research has shown participants were concerned about the burden of decision-making this might place on them (Hartman *et al.*, 2020; Steedman, Kennedy and Jones, 2020).

Regarding concerns about the burden these methods place on individual citizens, some have criticised these types of approaches as ideologically aligned with neoliberal positions. These arguments rest on the way in which individuals are expected to look after their own interests concerning how their data is used, and how these approaches draw responsibility away from the state or collective and place them on the individual (O'Hara, 2019; Krutzinna, 2021).

Despite the stated intention to move away from Silicon Valley based business practices, the AuroraAI project used very similar instruments for ethical oversight and algorithmic bias mitigation as can be found in these organisations, such as an ethics committee and a loose, adaptable structure for handling ethical concerns (Moss and Metcalf, 2020). This may suggest that working practices around algorithmic bias mitigation, and organisational ethics work more generally, are highly influenced by the structures of Silicon Valley business practices,

and that this also influences the professional imagination as to how these concerns may be handled.

While some participants were concerned about large private entities collecting and controlling large quantities of data, and therefore about the privacy of citizens, the AuroraAI ethics committee participant expressed concern that there did not seem to be a strong justification for collecting large quantities of citizens' data for the AuroraAI system. Some participants, however, thought this would provide the state with new insights into how to tackle social issues. A member of the AuroraAI ethics committee said:

> "[but] we've known this situation for twenty or thirty years, and it's always been a question of lacking in political will to allocate certain resources in certain regions and services and service sectors. [..] Building an entire AI system and building all these ecosystems and networks of people is not cheap. Will we be getting our money's worth?" (P5, AuroraAI ethics committee)

The gap between the project's senior management, the project's ethics committee, and the project's more technical workers' views around ethics work seemed to stem from both their professional priorities and duties on the project, as well as their differences in political and value-based assumptions. For example, one of the project's senior advisors most important ethical priorities was the malicious influence of Big Tech companies on the state and the AuroraAI project's potential for facilitating international cooperation rather than algorithmic bias. However, when discussing the values promoted by the AuroraAI system itself, some participants stated that they did not perceive the system to be promoting any particular set of values, as the 'life hacking' component of the project was open for citizens to input their own goals. In relation to this, one participant said:

> "It's not something that we as a government or any similar entity [say you should be aiming for] because we are not in a position to say that this is good life or this is not good life. So, the individual itself sets the goal and we are not setting our values into the system. This is value independent" (P3, AuroraAI senior management)

Despite senior management participants perceiving the 'life hacking' segment of the project as "value independent", much of the promotional material for AuroraAI mentions that the system aims to empower citizens in their health and career options. Additionally, this value independence was questioned in a document produced by the AuroraAI ethics committee, which stated "the program's fundamental values in this regard are not sufficiently opened up and made explicit" (D4). When asked what had been learned from the AuroraAI project's process which might be applicable to other public sector organisations looking to mitigate bias, one of the AuroraAI ethics committee participants said:

> "An AI system that will have direct influence and effect on people, real life, real people, will need to ask themselves […] 'What are the values that our work is based on, to make it sort of like transparent and make it visible first even to ourselves, so that we are completely aware of what we are trying to achieve and why.'" (P5, AuroraAI Ethics committee)

During this discussion, the AuroraAI ethics committee participant advocated a design process similar to Friedman *et al.*'s (2017) Value Sensitive Design (VSD) methods, which encourage developers to consider the values embedded in technologies throughout the development process. Discussion with participants seemed to reveal a mismatch between the ways in which the values of the project were conceived. Some participants perceived the technology as being flexible, not imposing any particular value system on citizens, while some on the ethics committee questioned "whether [AuroraAI was] underpinned by the notion of humans as rational operators who optimize their own attributes and opportunities and possibilities?" (D4). This observation echoes comments from Green (2021), who states that a lack of consensus around what can be considered 'the social good' can allow those in power to present their judgements as fulfilling normative values and thus being desirable (Green, 2021). Specifically, these observations suggested a lack of shared understanding regarding the values the technology aimed to promote, and potentially, allowed normative values regarding individuals' desire to pursue health or career orientated goals to be presented as neutral.

Operationalising Good Practice Concepts

Closely linked to the above theme is the way different practitioners sought to operationalise concepts such as discrimination, bias, transparency, and co-design. Below I expand on how practitioners utilised different understandings of these terms, and how this influenced attempts to operationalise good practice in algorithmic bias mitigation.

All participants discussed algorithmic bias in relation to the concept of discrimination, however, as a participant from an AI Ethics organisation said:

> "The word discrimination means quite different things to some machine learning engineers than it does to a lawyer." (P13, Ada Lovelace Institute)

Data practitioner participants on the AuroraAI project generally described discrimination as coming "downstream" from either biased data or algorithms. These participants often considered discrimination from a perspective of their legal responsibilities as practitioners, with the most discussed example being that of disability discrimination. This supports observations that the term 'discrimination' in machine learning, compared to the term 'unfairness', has come to take on its legal meaning and connotations (Edwards and Veale, 2017; Selbst *et al.*, 2018). In addition, viewing this concept through a legal lens, participants often described discrimination as either being an accessibility issue, or an extension of accessibility issues. This was due to my participants' understanding that they had a legal responsibility to ensure software was accessible. This contrasted with the way participants who were either AI Ethics experts or on the AuroraAI ethics committee used the word discrimination; to refer to both illegal and legal forms of prejudice against marginalised groups. For example, a legal form of prejudice which some participants mentioned being concerned about was perpetuating the digital divide between certain groups. This was due to participant understanding that some sections of the population were more digitally capable,

and thus providing tools and services to those who were already best able to make use of digital services ran the risk of widening the gap between these groups.

With regard to the term 'bias', participants across my interviews used it inconsistently. Even amongst my participants with a technical background, the word was used in a very loose manner. Sometimes it was used to describe a particular bias (selection biases, instrument biases, cognitive biases etc.), and sometimes it was used interchangeably with the word 'discrimination'. This compounded communication difficulties between participants from a technical background and those on the ethics committee.

In addition, one participant, who worked with young offenders, said that the prison service used scientific papers to inform their data modelling practices and data analysis. She described this as being particularly important, as without this knowledge it would be difficult to recognise potential biases within their data analysis process. This resembles what Jaton describes as " ground-truthing practices" (Jaton, 2021). Ground-truthing practices refer to processes whereby practitioners use referential evidence to guide data analysis. For example, when analysing whether a certain intervention assists young offenders' rehabilitation, scientific literature could be used to check for any known biases in these types of analysis. This therefore checks whether the patterns and interferences learnt during the analysis are those which resemble established characterisations of the phenomenon. Jaton positions these practices as providing moral pragmatism. However, afterwards, the participant reflected on the potential issues which might arise from this practice:

> "We have to take a critical look on the scientistic research on offending and offenders, and the criminological research on these issues, can we rely on the researcher results? Of course, science renews itself and recorrects itself in the process, when research develops, it's good to take into consideration this too, that maybe even scientific research is not always without bias" (P10, AuroraAI, Pilot Project Lead).

This suggests two things. Firstly, ground-truthing practices which rely on institutionalised knowledge production, such as scientific research, may be an important aspect in the shaping of practitioners' perception of biases, and may contribute to the proliferation of algorithmic bias. Secondly, attempts to provide precise definitions of algorithmic bias (Danks and London, 2017), may lack the necessary flexibility to encompass the broad spectrum of biases which my participants discussed. Regarding the diverse interpretations of the term 'algorithmic bias', an Al ethics expert participant commented:

> "It's not that we're looking for the perfect definition that's sort of the exactly perfectly worded one. It's about ensuring that we are creating these definitions that are meant to be flexible and encompassing, and incorporating different viewpoints and perspectives." (P13, Ada Lovelace Institute)

However, this need for terminological flexibility may sit in tension with traditional computing frameworks which require precise definitions for the logical processing present in algorithmic systems.

Regarding transparency, participants again had differing ideas as to what practicing transparency on the AuroraAI project might mean, and how effective these practices might be. The AuroraAI project had a strong focus on allowing public access to all design and planning documentation. These documents were uploaded to the project's Google Drive, which is publicly accessible through its Slack channel. The Slack channel also provided public access to anyone wanting to "get involved" in the project, and posted details about upcoming meetings which were open to anyone to attend. The level of transparency around the AuroraAI project was particularly impressive, and hardly seen in public sector AI projects.

However, there were concerns from AI Ethics expert participants that these mechanisms do not provide accessible forms of transparency to the general public; "it probably requires some level of expertise to understand those project documents, and I'm not sure that without more information they'll be particularly useful to anybody" (P11, Data Justice Lab). Moreover, one participant stated that these forms of transparency often miss key details:

> "the piece that we're finding people aren't making available [is] the impact of the systems that we looked at on service users and on how resources were used." (P9, Data Justice Lab)

These insights echo Ananny and Crawford's (2018) comments that transparency without consideration of the intended audience can obfuscate, rather than illuminate, details available to citizens about what a project entails (Ananny and Crawford, 2018). Bates *et al.*'s (2023), writing on the concept of transparency from their work with public sector research partners in the Living With Data project, argue transparency practices need to be socially meaningful. Socially meaningful transparency practices are ones which "[foreground] the needs and interests of those who require information to be transparent for them to understand data-based systems," in contrast to those which "[centre] the interests of data-systems developers or others who are engaged in transparency for the purpose of compliance or public relations" (Bates *et al.*, 2023, p2). While the AuroraAI project's transparency practices are very extensive, it is uncertain whether they are socially meaningful. In the context of the AuroraAI programme, the way in which transparency was operationalised focused more on engaging a wide community of practitioners, in addition to creating an international stage for AuroraAI as an example of ethical AI, and so may not be meaningful to actors most interested in the risk of algorithmic bias posed by the technology.

Although, many of the participants from the AI Ethics group were still very positive about AuroraAI's transparency practices, some AI Ethics participants thought these transparency practices needed an accountability mechanism incorporated alongside them. One such method which was suggested was AIAs (algorithmic impact assessments), which are designed to build on previous impact assessments in other sectors, such as human rights, environmental, and data protection impact assessments. Still, little is known about how these mechanisms might work regarding algorithmic technology development. One participant said:

> "In practice, algorithmic impact assessments haven't been done much, there aren't many case studies, there aren't many what does this look like on the grounds" (P8, Ada Lovelace)

Lastly, the AuroraAI programme's documentation mentioned the use of co-design principles as part of the design process. When asked about these, participants referred to the methods used by certain pilot projects, and to a survey of Finnish student councils. However, when asked further questions about these pilots, the methods used did not seem to be fully co-design, but more closely resembled the collection of user feedback. For example, one participant working as a researcher on the project said that one of the project's preliminary trials was run by the church, and focused on young people using AuroraAI. The process which the participant described involved the church giving the young people a survey as part of the pilot, although little was known about how the survey results fed back into the pilot. When this participant was asked, they responded that they trusted the church to act fairly in these circumstances. This confirmed that the practices did not follow those of co-design, but instead solicited user feedback as part of the pilot. Furthermore, these pilots did not elicit feedback from the citizens most at risk of algorithmic harm, but from privileged groups which were easier to access. One participant said:

> "[the] young people who are elected – they have these elections in schools, so the young people themselves elect their representatives to these youth councils. And they tend to be, for the most part, ambitious, hardworking, doing well at school […] and they're supposed to represent all the not so well-to-dos and school dropouts in the area" (P5, AuroraAI Ethics Board).

This also suggests that in common with the way transparency was operationalised on the project, 'co-design' was operationalised in a manner which did not serve to substantially strengthen users' input into the design process, nor to specifically consult marginalised groups as part of this process. it could be suggested that by primarily soliciting the input of less marginalised students to speak for all students, this mechanism may only further reinforce current power relations between these groups – instead of providing marginalised groups an opportunity to contribute to the design process. Through talking with participants this seemed, at least in part, to be due to the ease with which these systems could be put in place compared to the more challenging work of engaging with marginalised communities.

Communication and project planning

As discussed above, participants had very different ideas of what constituted good practice, and alongside this, also operationalised many of the key words in their working vocabulary quite differently. This section explores how this lack of shared understanding, as well as project management styles and competing project pressures, allowed for concerns about algorithmic bias to be sidelined.

In the initial stages of the project, an AuroraAI ethics board participant recalled there was some disagreement between the sponsoring stakeholders about what the project's goals

should be, and how to go forward with them. One of these disagreements centred on how the project's scope had expanded since its initial conception. Originally, the project was intended to be a simple recommendation system – users would type in an issue they needed help with, such as getting a divorce, and the system would direct them to the most appropriate public and private services in their area. However, as the project developed, the idea expanded so that AuroraAI would provide more expansive life advice. It would also provide recommendations as to how to retrain for work, and it would invite users to customise career and health goals, similar to the way people use a Fitbit. One participant stated:

> "Imagine if in addition to the Fitbit you would have a – well, I don't know, your medical records, whatever data there is about you in the government systems helping Fitbit to make you feel even better" (P1, AuroraAI)

According to a participant from the AuroraAI ethics committee, some of the project's original private sponsors were put off by this change of direction, concerned that this would be a much riskier endeavour than the project's original goals. Tension over how to assess the risks associated with AuroraAI remained during the development of the project, and challenged attempts to address algorithmic bias. One participant, a researcher on the project, commented that the risks of algorithmic bias would be minimal if they supplied users with enough control regarding what data the recommender system had access to. Consequently, this participant saw the development of the 'digital twin' aspect of AuroraAI to be highly important with respect to issues of bias. This was in part due to the participant's belief that by providing users with control of their data through their 'digital twin,' no one would be denied or refused a service on the basis of their data or any data processing. Rather, they would be able to make up their own minds about these matters and adjust their data preferences accordingly.

These participants framed their understanding of bias using normative expectations of citizens shouldering individual responsibility for their data choices. Returning to the issue of 'digital twins' and individual responsibility discussed earlier, these approaches assume individuals are best placed to protect themselves against discrimination, and that sharing (or not sharing) their data will assist individuals in this goal. However, approaches such as these require citizens to be aware, prior to sharing their data, how their data may be used to discriminate against them. One senior tech participant on the project stated that they were not concerned about bias due to the control they were giving citizens over their data – and that citizens would be able to lie about their data if they wanted to. However, this presumes the user of the AuroraAI system can predict when an algorithm may be able to discriminate against them before the fact. As discussed earlier, while there is public interest in personal data store approaches (Hartman *et al.*, 2020), there is both academic and public concern that these place an unfair burden on citizens to manage their data (O'Hara, 2019; Steedman, Kennedy and Jones, 2020). Furthermore, these approaches may individualise what are collective struggles – that is, people are discriminated against due to belonging to a marginalised group. For that reason, while these approaches may provide a mechanism to divest corporations of power, it is unclear how they might address issues of algorithmic bias.

The AuroraAI team seemed to have difficulties resolving differences in perceptions of the risk level associated with their project. Both participants who worked with data on the project, and the ethics committee participant, brought up examples in which they suspected the "other side" had misunderstood the level of risk associated with the technology they were trying to develop. Whereas participants who worked with data considered the project to 'simply offer recommendations', the ethics board participant was more concerned that any recommendations the system produced could be perceived as coming from the Finnish state. Therefore, users of the system could feel pressured into following the system's recommendations if they perceived the recommendations as being official government advice. While reflecting on the moral issues underlying recommendation systems, the ethics board participant commented:

> "there's also the power imbalance, some people may perceive it as, 'Oh my god, the artificial intelligence of the state of Finland is now telling me to contact this service or use that service, or start jogging […]' and they will go like, 'Okay.'" (P5, AuroraAI Ethics board)

AI Expert participants shared similar concerns about the potential for moral concern regarding recommender systems. They did not see these systems as neutral. However, these participants believed that compared to algorithmic technologies which allocated welfare resources, a recommender system seemed to be of less moral concern.

Additionally, participants with data-focused roles on the project sometimes found it difficult to understand the recommendations from the ethics committee. Participants mentioned that they wanted more practical suggestions from the ethics board. A participant who worked with data said he often felt confused about how to interpret the feedback the ethics team supplied. He described a situation in which implementing the feedback would require the AuroraAI system to be able to do two things which are almost impossible simultaneously; to be aware of a users' protected characteristics when making recommendations, to avoid providing disabled people with unsuitable recommendations, while at the same time not storing any data which could reveal those characteristics. He said he understood the concern, but found the lack of knowledge around how these systems worked frustrating:

> "some people know more about AI and machine learning and how the practical implementation now is being implemented and some people have zero information on that." (P5, AuroraAI data engineer)

This suggests that despite a willingness amongst participants to work together to mitigate algorithmic bias, efforts were often hampered by a lack of shared understanding about the potential risks and capabilities of the technology in question.

As discussed in Chapter 4, Metcalf *et al.* (2021) suggest that impact assessments can act as boundary objects in algorithmic bias mitigation. Following from this, I would argue that a project's ethical procedures, such as an ethics committee and the working practices which embed the committee into the wider project, can additionally be conceived of as a boundary object. As noted above, boundary objects are objects which are understood locally within

disciplines but are malleable enough to be understood across contexts (*ibid*). As such, experts are able to use these objects to facilitate collaboration and co-construct their understanding of the potential impacts and harms of a project (Metcalf, *et al.*, 2021). Carlile (2002) writes that when practitioners are working across knowledge boundaries, difficulties arise through a lack of shared systems for communication and problem solving. Each community of practice has developed their current knowledge base to address the particular problems to which their work is orientated Or, to put it another way, each community has different "stakes" – the expectations and deliverables towards which their knowledge and skills are utilised when co-operating on a project (Carlile, 2002).

The AuroraAI participants expressed a willingness to engage in work across their disciplinary boundaries, and to try and create new processes for cooperation on the project. However, the participant from the ethics committee often described this as a struggle, and felt it was not uncommon for the ethics committee's concerns to be sidelined. As mentioned earlier, a senior project advisor stated that ethics should not be a "rock in the shoe" (P2), meaning that some participants believed that the ethics committee should not impede the progress of the project. In other words, it was preferable that ethics work did not impede what was 'at stake' for other types of practitioners on the project. This complicates the expectation that working with boundary objects primarily involves finding ways for different practitioners to communicate and work together, and supports the argument that this type of work is a political activity, where there is a risk of practitioners using boundary resources to further their own interests (Kimble, Grenier and Goglio-Primard, 2010; Grenier, 2006). While the project's leadership agreed to the need for an ethics committee and committed to embedding the committee within the project infrastructure, there was little in the way of mechanisms which created accountability between the ethics committee and other stakeholders on the project.

As discussed earlier, the speed of the project's development often meant key decisions were made prior to input from the ethics committee. An AI Ethics expert participant from the Data Justice Lab stressed that a slower and more cautious approach was less likely to cause issues, "when things go wrong what we've seen is because things were rushed and they were put in place too soon" (Participant 9, Data Justice Lab). This was a tension in the development practices on the AuroraAI project, with participants often describing agile working environments where participants who focused on data work were more concerned with ensuring the technical aspects of a project were working. This difference in working practices in the two groups seemed to cause friction regarding how the project could go forward, with the ethics board often feeling left behind.

Challenges in the project's wider social environment

The previous section discussed how communication and project management practices led to moral concerns being sidelined. In this section, I will discuss how the project's wider social environment presented further challenges in addressing algorithmic bias.

A reoccurring theme across the participants who worked on AuroraAI and those who worked in AI ethics organisations, was the influence of funding structures on the feasibility of implementing good practice procedures. Participants said that project funding could limit how much work could be done to ensure good practice was followed. This might be due, for example, to funding not covering extensive rounds of public consultation. Moreover, the AI ethics participants also explained that when technologies are funded by either state funds or venture capitalists, they are generally funded with conditions or expectations attached to them. A participant from the Data Justice Lab explained how funders, or the origins of the funding, often set the framing of the societal 'problem' and of the potential 'solution' – and the solution framed by funders may reinforce certain worldviews and value assumptions. Drawing on the example of the police transformation fund in the UK, a participant from the Data Justice Lab described the following issue:

> "[funding technologies is] seen as a form of action and seen as tackling crime etc. But it's an investment question. You know, is this the right way to invest resources into that area? […] we're sort of just introducing technology as a way to be seen to be more efficient with resources, but not actually necessarily with any evidence that they actually do what they are meant to do." (P11, Data Justice Lab).

On the AuroraAI project, participants engaged with data work saw other issues in relation to funding. Primarily, they noted that the expectations attached to a project's funding, and thus what the available time on a project is dedicated to try to achieve, may not actually be technologically feasible – "it simply sounds impressive" (P4). To put it another way, the original project goals might not be possible, but simply sounded good in a funding bid. One participant explained that in his opinion, they were not really working on anything like 'AI' on the project, but due to the current AI hype, it was beneficial to call it AI to give an impression of a project being "cutting edge" (P7, AI Ethics expert).

> "if you use the term AI and machine learning, you are able to get people's attention. [...] It gives an impression that you are ahead, you are a pioneer, you are doing something that not everybody can, […], you are a forerunner" (P4, AuroraAI, Data Pracitioner)

This expansion of what the word "AI" means added to the tensions experienced by the technically and ethically focused participants when attempting to assess and communicate the project's potential harms. In one situation, one AuroraAI participant in a data focused role was concerned that this led some ethics experts to over-exaggerate the potential harms posed by algorithmic technologies, claiming that "their background in AI is coming from *Terminator 3*, the movie and things like that" (P4, AuroraAI, Data Practitioner). Indeed, during the course of my interviews, many participants would use science fiction as touchstones to communicate their concerns about the project – including *Big Brother*, George Orwell's *1984*, and the Three Laws of Robotics mentioned in Isaac Asimov's *I, Robot*[6]. The ethics committee

[6] 1) A robot may not injure a human being or, through inaction, allow a human being to come to harm. 2) A robot must obey orders given it by human beings except where such orders would conflict with the First Law.

participant mentioned the predictive system in the film *Minority Report*, primarily due to the similarities he saw in relation to data collection and predictive analytics by the state. However, while these touchstones were used as communication tools, the AuroraAI project was very far removed from the technology presented in these fictional worlds. This suggests some practitioners' perception that the image of AI which exists in the collective imagination, one influenced by science fiction and technological hype, may cause difficulties for practitioners when attempting to assess and communicate the risks of these technologies.

While the shared imagination around AI, and the hype which accompanied it, was described as being beneficial from a funding perspective, some participants felt it muddied the waters for realistic communication regarding the technology. This suggests the need for careful discussion around the actual threats presented by these technologies, and the need for practitioners to avoid making assumptions about other practitioners' perceptions of algorithmic technologies.

5.6. Discussion and conclusion

In this paper, I investigated the following research questions: What might 'good practice' on an algorithmic project look like? What challenges does good practice on an 'ethical AI' project face in practice? This was explored through interviewing practitioners on the AuroraAI project by the Finnish Ministry of Finance, as well as practitioners outside of the AuroraAI project who worked in an AI Ethics capacity. The results of this study suggest two key findings. First, that even in this purportedly progressive project, and within AI Ethics more generally, there is a lot of disagreement about what constitutes good practice in mitigating algorithmic bias and the types of solutions that may be practically implementable. This disagreement was further complicated by differences in understanding how key terminology should be operationalised, which was found in both participants from the AuroraAI group and the AI Ethics group. Second, the findings suggest that project management styles which focus on technological pursuits may not give enough time to focus on how to mitigate the impact of biases, and additionally, may allow moral concerns to be sidelined in favour of prevailing ideological assumptions. I expand on both below.

Regarding the first finding, participants often disagreed on how they should implement algorithmic bias mitigation approaches. While all participants on the AuroraAI project said they were committed to developing an algorithmic technology which improved society, challenges arose due to difficulties in conceptualizing ethical concerns and risks within their shared imagination. Participants operationalized different understandings of risk, with technical participants focusing on accessibility or legal frameworks, which influenced attempts to communicate about algorithmic bias during the course of the project. These differences often related to what was "at stake" for the participant in question (Carlile, 2002), with participants adhering to their professions' understandings of key terminology, as well as operationalising concepts in a manner which suited the deliverables expected of their role on the project. For example, operationalising co-design principles in a manner which more

3) A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

closely resembled user feedback, rather than actively engaging with citizens and empowering them through participatory mechanisms (Arnstein, 1969), given the ease of eliciting public opinion in this manner. Additionally, AuroraAI participants disagreed on the values which the project promoted. The AuroraAI ethics committee perceived the project as potentially promoting values around individual autonomy and self-interest, whereas senior management on the project perceived the system to be value free. However, at the time of data collection, this issue had not been resolved, as this had been an early design decision prior to establishing an ethics committee on the AuroraAI project.

This highlights the importance of recognising how organisational dynamics influence the use of boundary objects. Not only do practitioners need to create a shared understanding of the practical challenges faced when working across disciplines, but they must also find ways of overcoming conflicting personal and organisational value assumptions during this work. This supports Green's position that in the absence of a strong definition of the 'social good', those in power may use normative assumptions to present their values as fulfilling the 'social good' (Green, 2021). In the case of the AuroraAI project, normative assumptions regarding the importance of individual responsibility and self-improvement were relied upon to justify the aims of the project. Moreover, my findings suggest that these conflicts are further complicated in a public sector context, where the state's value assumptions will influence the values seen as desirable during the project.

Additionally, the findings presented above highlight two ways in which financial incentive models, in both public and private technological development, present particular challenges when attempting to mitigate bias. Firstly, AI Ethics participants were concerned about how funders set the criteria and expectations for what a project adheres to, and thus the 'problem-solution' framing to the technology prior to its development. Secondly, funding availability changes what design choices can be materially conceivable, such as how often the public can be consulted, in what way, and what post project feedback systems look like. AI Ethics expert participants, particularly those from the Data Justice Lab, positioned solving the 'problem-solution' funding issue as having greater importance than creating organisational approaches to mitigating algorithmic bias. Using a social justice-based framing of algorithmic bias, these participants often saw good practice in this context as practices which involved reflection regarding the wider system within which these technologies are embedded. In contrast, AI Ethics expert participants from the Ada Lovelace Institute were more concerned about how limited funding could restrict a project's ability to consult the public and implement post project feedback systems. This highlights the range of approaches which AI Experts consider to be a priority when seeking to mitigate algorithmic bias.

Furthermore, participants from different professional backgrounds utilised different framings of the ethical issues they encountered when developing algorithmic technologies, with the technical participants relying more on algorithmic and data framings of the issues they were engaged with during the course of their work (Selbst *et al.*, 2018). These differences in understanding, combined with the limitations presented by computing systems, made communication about bias between practitioners difficult. These issues were further compounded by the hype around algorithmic technologies, which created difficulties for

practitioners when communicating concerns about the potential consequences associated with the project. This study suggests that efforts to mitigate bias can be hampered by a lack of a shared understanding of core aspects of a project, and confound attempts to communicate regarding the implementation of algorithmic bias mitigation methods. This means that effort is required to create a shared understanding of core concepts across stakeholders with often disparate skillsets, expertise, knowledge-bases, values, and beliefs.

Regarding the second contribution, it was found that project management styles which focus on technological pursuits may not give enough time to focus on how to mitigate the impact of biases. Participants on the AuroraAI project struggled to find the right moment to consider bias during the project lifecycle, which was often difficult due to competing demands with regard to the priorities set by the project's senior advisors.

The project relied on an ethics committee as its primary instrument for addressing algorithmic bias. However, at the time of this research, it was still uncertain how the ethics committee would be embedded within the project's working practices. Despite project management's intentions that the AuroraAI project should resist the types of data practices found in large technology corporations, the ethics committee on the AuroraAI project struggled with many of the same difficulties found in Moss and Metcalf's study of Silicon Valley ethics workers (Moss and Metcalf, 2020). This included tensions between team members regarding the purpose of the ethics committee, as well as a desire for the ethics committee to act 'constructively' on a project, so as not to hamper other practitioners' deliverables. This is further complicated by the typical agile working practices found in technological development, which encourage fast-paced progress and early tangible deliverables, and which does not provide the time necessary to consider the potential ethical consequences arising from an algorithmic project.

However, whilst there were challenges regarding AuroraAI's implementation of an ethics committee into the project, some of the project's ethical practice was very encouraging. The level of transparency on the AuroraAI project, allowing citizens to access meeting documents, the project Slack channel, and attending meetings themselves if they were interested, is hardly seen in public sector projects. Still, AI Ethics participants saw there was room for improvement in their transparency practices, specifically by ensuring that their transparency practices were meaningful, and by linking these to accountability mechanisms such as AIAs (algorithmic impact assessments). Bates *et al.* (2023) argue that transparency practices need to be socially meaningful, by "[foregrounding] the needs and interests of those who require information to be transparent for them to understand data-based systems" (Bates *et al.*, 2023, p2). To create socially meaningful transparency practices, Bates *et al.* (2023) argue that transparency practices should be a collaborative process between developers, citizens, third sector organisations, and experts. On the other hand, the Ada Lovelace Institute is in the process of developing an AIA for use in the UK in an NHS context. AI Ethics participants explained that assessments such as these are usually linked to accountability structures, and would provide a level of oversight to the way the risks of a project are conceptualised and managed.

Interwoven within these challenges, there is the ever-present difficulty within innovation and technological development of hype and techno-utopianism. Large technology companies such as Apple, Facebook, etc., have a strong influence on what is seen as possible, both in terms of project aspirations and working practices, by governments and designers, and their presence in the global sociotechnical imaginary remains deeply influential.

Furthermore, this paper contributes towards answering the overarching research aims of my thesis: to investigate how DWP practitioners might mitigate the impacts of algorithmic bias. The findings presented in this study suggest that while there was disagreement regarding good algorithmic bias mitigation practices, AI ethics experts suggest a range of approaches. Of note in this paper are the AuroraAI ethics board members' suggestion of VSD, the Ada Lovelace Institute's suggestion of Algorithmic Impact Assessments, and the AI Ethics experts' combined suggestion of improved critical thinking about data, technology, and the relationship between algorithmic technologies and the wider environment of social inequality and injustice. Moreover, the findings in this paper indicate that the approach chosen is not the only aspect of importance, but that the way it is integrated into the project is equally important. While the AuroraAI project utilised an ethics committee as its primary instrument for algorithmic bias mitigation, its lack of integration into the project's working practices presented challenges regarding its effectiveness. In the next chapter (Chapter 6), I investigate how these insights might be adopted within a DWP context.

Chapter 6: Paper three, The influence of DWP organisational culture on the adoption of algorithmic bias mitigation practices and implications for practice

In the previous two chapters, I presented the first and second of my three thesis papers. The first paper (Chapter 4) investigated DWP's current working practices regarding algorithmic bias mitigation practices. In the second paper (Chapter 5), I explored how another public sector organisation was approaching algorithmic bias mitigation, and what might be learned from their working practices.

Following from these papers, I return to the overall research aim of my thesis: to investigate how the DWP might mitigate the risks of algorithmic bias. Thus far, my findings suggest two things. First, the importance of the organisational context and working practices in the adoption of algorithmic bias mitigation practices. This was indicated in my first research paper, as it was found that participants strongly relied on legal frameworks due to their position as civil servants, and previous research about conditionality within the department had a strong influence on the framing of subsequent research. In my second research paper, this was indicated by the difficulties practitioners had in actioning algorithmic bias mitigation methods, because of the way in which the ethics committee was embedded within the AuroraAI project. The second point suggested by my research so far is that there is a lot of disagreement about what constitutes good practice in mitigating algorithmic bias, and the types of solutions that might be practically implementable. Despite these difficulties, participants in my second paper (Chapter 5) suggested the following approaches to create a more unified approach in mitigating algorithmic bias: Value Sensitive Design, Algorithmic Impact Assessments, and improved critical thinking about data, technology, and the relationship between algorithmic technologies and the wider environment of social inequality and injustice. In this third and final paper of my thesis, which is presented in this chapter, I focus on the following research questions: how can the insights from the findings of paper two be integrated into a DWP context?

6.1. Introduction to paper three

Algorithmic bias mitigation efforts that focus on debiasing either the datasets or the algorithms themselves have been well documented. For example, debiasing techniques have been developed where data practitioners either remove or alter the protected characteristics of individuals in the dataset, to make it more difficult for algorithms to make inferences on the basis of these characteristics (Ghadiri, Samadi and Vempala, 2021). However, so far socio-technical methods for mitigating these harms have received less attention (Ada Lovelace Institute, 2022). One such prominent method is algorithmic impact assessments (AIAs), which were proposed by the AI Now institute in 2018 (Reisman *et al.*, 2018). Other techniques expand on both Cathy O'Neil's assertion that models are "opinions embedded in

mathematics" and Friedman's (2017) work on Value Sensitive Design by focusing on the values embedded within these technologies. These methods explore how to assess and leverage the processes which lead values to become embedded within technologies (O'Neil, 2017: Friedman, Hendry and Borning, 2017: Umbrello and van de Poel, 2021: Lüthi, Matt and Myrach, 2021). As these practices are in their infancy, there are questions to be answered regarding their effectiveness in mitigating algorithmic bias. Furthermore, there is little research about how organisations might adopt these methods, and how they might be sustained through institutional practice.

This paper addresses this gap by investigating how public sector workers might implement socio-technical algorithmic bias mitigation methods, and what the barriers and opportunities are within this space. This paper presents three findings regarding algorithmic bias mitigation. The first is that it is difficult for public sector data practitioners to align technologies to the social justice values that underpin socio-technical algorithmic bias mitigation techniques when servicing a large diverse public. The second finding is that practitioners perceive there to be a lack of clarity in organisational guidance and legislation regarding fairness and discrimination. I suggest this can lead to additional uncertainty concerning how conflicting needs within the population should be addressed. The third finding is that participants perceived workforce diversity as important to algorithmic bias mitigation efforts. The findings of this paper are derived from data collected from conducting a series of seven educational workshops on algorithmic bias mitigation, and seven follow up interviews with practitioners in The Department of Work and Pensions (DWP). The workshops focused on how algorithmic bias might develop, and explored socio-technical bias mitigation approaches, such as algorithmic impact assessments and value sensitive design. After these workshops, participants were invited to take part in a follow up interview, to allow them to reflect on the content of the workshops and its relevance to their working practices.

The paper is laid out as follows. In the first section, as part of a literature review, I examine the literature on practical algorithmic bias mitigation mechanisms such as algorithmic impact assessments and value sensitive design. I then discuss theories of organisational change, and how organisational change has been attempted in the related area of diversity initiatives. In the third section, I detail my methodological approach and data collection process, and explain how these are suited to investigating this problem. In the fourth section, I discuss my findings, how they relate to the literature on socio-technical bias mitigation efforts, and what this may mean for the civil service organisational context. Finally, I discuss these findings in relation to the wider literature and offer potential recommendations.

6.2. Literature Review

In this section, I examine the literature on practical algorithmic bias mitigation mechanisms, such as algorithmic impact assessments and value sensitive design. I subsequently, I discuss theories of organisational change, and how organisational change has been attempted in the related area of diversity initiatives.

Algorithmic Bias Mitigation methods

In this section, I discuss socio-technical algorithmic bias mitigation methods. In previous chapters, I have outlined the limitations of technical framings of algorithmic bias. Because of these issues, data justice organisations and academics have been pressing for the adoption of algorithmic bias mitigation methods which go beyond algorithmic and data-based framings. Two prominent bias mitigation methods which have been discussed include the use of algorithmic impact assessments and value sensitive design (Friedman, Hendry and Borning, 2017; Lüthi, Matt and Myrach, 2021). These practices focus on design exercises and bureaucratic processes to foster a space in which data workers can consider how algorithmic bias might develop on an algorithmic project. Moreover, these practices aim to assist practitioners in embedding 'user' insights into the design process. In addition to these approaches, it has been suggested that data practitioners need to develop more critical thinking regarding data work (Ferryman and Pitcan, 2018; Green, 2020). I describe the above in detail below.

Algorithmic impact assessments (AIAs) have gained traction as an algorithmic bias mitigation method in recent years. They have been recommended by the AI Now Institute, a research institute which focuses on social and policy research about artificial intelligence, to assist in mitigating the risks of algorithmic bias (Reisman *et al.*, 2018). They have also been recommended by the Data and Society research institute, who do similar work on the social implications of algorithmic technologies (Moss *et al.*, 2021). AIAs build on standard practice in other sectors such as environmental protection, data privacy, and pharmaceutical risk management (Reisman *et al.*, 2018; Selbst *et al.*, 2018; ICO, 2019). They are intended to provide a framework for organisations to assess the risks associated with their technologies, as well as facilitating wider discussion of the fairness of systems. Additionally, the development of AIAs responds to calls for increased mechanisms to enable citizens to better understand the impacts of algorithmic technologies, and to provide the public with the information required to hold organisations accountable for their decisions regarding algorithmic technologies (Binns, 2018). However, it is worth noting that the benefits regarding public engagement are only possible if AIAs are released publicly. Currently, the only AIA which has been legally implemented is the Canadian AIA enforced by The Government of Canada's Directive on Automated Decision-Making (Ada Lovelace Institute, 2022). The AIA of The Government of Canada includes a 60-question questionnaire, designed to make developers think reflexively about their design choices, and to consider whether certain choices (such as using sensitive data, having opaque models, etc.) are necessary (*ibid*). Organisations are then required to upload their completed AIA to The Government of Canada's Open Government Portal prior to deployment of their algorithms (Treasury Board of Canada Secretariat, 2021).

In a UK context, The Ada Lovelace Institute recently released a case study of the AIA process they have been developing in collaboration with the NHS (Ada Lovelace Institute, 2022). This case study focuses on the NHS AI Lab's National Medical Imagining Platform (NMIP), an initiative which draws medical-imaging data from across NHS sites. The NMIP project aims to allow companies and research groups to use the data to develop AI models which can detect

different clinical conditions, assisting in faster diagnosis times for critical conditions such as cancer (Ada Lovelace Institute, 2022).

In the NHS case study, the process involves users of the data completing a reflexive questionnaire which asks the data applicant questions regarding their plans to ensure consideration of ethical challenges. Once these have been completed, initial applications are filtered through a Data Access Committee (DAC) embedded within the NHS, which would constitute doctors, patients, and data experts as part of the committee. The creation of the DAC aims to link the process to accountable structures within the NHS context. Then, once the data applicant has submitted their initial AIA, the committee would coordinate participatory workshops with citizens to understand citizen concerns around the potential technology. Once the participatory workshops have been conducted, the data applicant team would have a chance to reiterate their application based on the workshop attendees' feedback, prior to a final decision by the committee (Ada Lovelace Institute, 2022). Presently, the Ada Lovelace Institute are in the process of piloting the proposed process and collecting data regarding the effectiveness of the process.

While proposed AIA processes differ, many of the core ideas remain the same. They aim to provide organisations with the space, and necessary questions, to identify and assess the potential algorithmic bias risks associated with embarking on an algorithmic project. Furthermore, they aim to support transparency regarding the design process of algorithmic technologies (Reisman *et al.*, 2018). However, for AIAs to provide these benefits, it is critical that they are linked to accountability processes, including a legal basis which requires truthful answers to the auditors of AIAs from those who are audited (Loi and Spielkamp, 2021).

However, there is concern that AIAs may hamper understanding of the very biases they are designed to mitigate. Metcalf *et al.* (2020) suggest that AIAs risk creating organisational metrics which inappropriately distance themselves from the potential harm caused by algorithmic technologies (Metcalf, *et al.*, 2021). They argue that unless those at risk of the impact of these technologies are consulted during the process of an AIA, then the impacts recorded as part of this process may not accurately represent the proposed technologies' potential for harm. Moreover, Metcalf *et al.* (2021) question whose expertise should be used to construct the impacts (or harms) documented within AIAs (Metcalf *et al.*, 2021). This follows on from research in other contexts which suggests that the focus of assessment instruments on risk assessment has led to harms being underplayed for social or political reasons (Guthman and Brown, 2016; Murphy, 2001). For example, these may include government agencies prioritising their relations with commercial organisations over citizens' interests (*ibid*). In another example, in an analysis of risk assessments conducted on toxic pesticides in California's strawberry farming industry, Guthman and Brown (2016) suggest the US Environmental Protection Agency and California's Department of Pesticide Regulation prioritised evidence submitted by industry when conducting a risk assessment on the safety of the buffer zone, which dictated the size between areas where pesticides could be used and residential areas. In doing so, they prioritised this evidence over evidence activists provided on the harmful effects of these pesticides (Guthman and Brown, 2016). This suggests that the

social structures surrounding AIAs influence the effectiveness of these mechanisms, and require researchers and practitioners' careful scrutiny.

In addition to assessment based algorithmic bias mitigation approaches, there have been calls to consider how values become embedded within algorithmic technologies, as a form of critical thinking about the political character of these technologies (Selbst *et al.*, 2018; Green, 2020, 2021). Approaches of this type draw on theoretical insights from STS (Science and Technology Studies, a field which aims to analyse the development of science and technology within its historical and political context) amongst other disciplines (Haraway, 1988; Jasanoff and Kim, 2009). Furthermore, insights from STS have formed the basis of well-established frameworks for developing socially sensitive technology, such as Value Sensitive Design (VSD), which was developed by Batya Friedman and focuses on how technologies can amplify or deprioritise certain values within the environment within which they are embedded (Friedman, Hendry and Borning, 2017).

VSD provides practitioners with tools for working with values during the design process, including conceptual, empirical, and technical exercises (Friedman, Hendry and Borning, 2017). Conceptual exercises focus on designers completing stakeholder analysis exercises, which aim to create a sense of understanding around different stakeholders' values within the proposed deployment context of the technology in question (*ibid*). Exercises involve the utilisation of research methods to investigate users' actual values regarding the proposed technology (*ibid*). Lastly, technical exercises involve the development of the proposed technology in conjunction with the information gathered during the previous two exercise stages (*ibid*). In following these steps, the VSD process aims to provide designers with a method for thinking about and designing technologies to meet the value-needs of the stakeholders involved in or impacted by these technologies.

In addition to AIA and VSD based approaches, there have been calls for data scientists to be more critical in how they use, interpret, and develop algorithmic and data-driven technologies. As discussed in Chapter 2, bias can enter a system due to the biases in a dataset and data collection practices, such as those associated with the use of proxy data, binary categorisation systems, and feedback loops (Eubanks, 2018; Noble, 2018; D'Ignazio and Klein, 2020). To counter this, it has been suggested that data scientists need to acknowledge the contextual nature of data, and move away from framings which position data as objective and neutral (Selbst *et al.*, 2018; Green, 2020). Green describes a contextual approach to data and algorithms as one which critically questions "the social relations, activities, and histories that shape any particular setting" (Green, 2020, p9).

Alongside these suggestions for data scientists to consider the wider context in which data is produced, academics have suggested the importance of considering what types of expertise and knowledge are required to recognise bias. Ferryman *et al.*'s (2018), researching biases in health data, suggest data scientists need to develop greater 'data empathy' to mitigate harmful biases in their working practices. Data empathy can be understood as understanding the origins of a dataset, what data collection practices produced the dataset, and the biases typical of a dataset of that type (Faghmous and Kumar, 2014; Mangal, Rajesh and Misra,

2020). Without sufficient domain and contextual knowledge regarding the data, "[there can be] a distance between these analysts and the data, specifically their lack of knowledge and direct experience of how, why, and where health data were collected" (Ferryman, Kadija, Pitcan, Mikaela, 2018). Furthermore, they argue "[t]his 'lack of data empathy' can limit their ability to recognize bias and optimize the analyses because they are too far 'from the source'". It is suggested that increasing data scientists' awareness of the contextual characteristics of a dataset will create working practices which recognise the limitations of algorithmic technologies.

However, despite substantial theoretical interest in this area, not much is known about how these methods may work in practice to mitigate the risks of algorithmic bias. In the following section, I review the literature on organisational change, which is relevant in order to address the current lack of research regarding how organisations might implement and sustain algorithmic bias mitigation methods, such as AIAs and VSD, through their working practices. Currently, there is little research on organisational change regarding algorithmic bias mitigation practices. Therefore, the following section discusses organisational change in relation to diversity initiative practices which focus on developing a more diverse workforce within an organisation, as these examples overlap with algorithmic bias mitigation in terms of their overarching aims as they are both working towards non-discriminatory working practices.

Organisational theory

Organizational change is pivotal to the issue of algorithmic bias mitigation. The algorithmic bias mitigation methods discussed and proposed by academics and independent research bodies require organisations to be able to implement them within their own working context, and in a way which is effective. It is not a particularly rare occurrence for organisations to receive criticism that new initiatives have done little to improve the situation they were designed to make progress in, particularly in the case of diversity initiatives across a wide array of sectors (Dobbin and Kalev, 2016). While studies such as those undertaken by Orr and Davies (2020), Veale (2017), and Holstein *et al.* (2019), have interviewed practitioners to understand how they are situated within the development of algorithmic bias, and what their responsibilities are within this space. Their analysis is primarily focused on how individual actors are constrained within their working context (Orr & Davis, 2020; Veale, 2017; Holstein *et al.*, 2019). The way in which algorithmic bias mitigation can be approached from an organizational or project perspective has received less attention.

To understand how the working context might influence the adoption of algorithmic bias mitigation methods, it is important to consider this issue from an organisational perspective. Organisations can be understood as comprising of three layers; the socio-psychological level, the organisational structural level, and the ecological level (Scott and Davies, 2006). The socio-psychological level focuses on the attitudes and behaviours of individuals within an organisation. The organizational structural level focuses on "structural features and social processes that characterize organizations and their subdivision", such as how the organisation

breaks down work, communication, departments, and authority within the organisation. Lastly, the ecological level, focuses on how the organisation as an entity operates within a larger system, such as in relation to other organisations or wider social or political structures (Scott and Davies, 2006). The ecological level may include wider communities of which individuals may be part, including communities of practice around their profession (Scott and Davies, 2006, p122). These layers are not self-contained, and the interplay between them is complicated and can be difficult to pinpoint in practice. For organisational change to be effective, new practices have to be adopted across these three layers, as well as throughout the differing levels of authority contained in the organisation. While it is sometimes understood that organisational change is managed by those in more senior positions, newer theories propose that the development of successful organisational change is developing at a middle management or grassroots level (Baker, French and Ali, 2021; Buchter, 2021).

Organisational change aiming to develop stronger ethical working practices has been especially fraught. Organisational change around diversity practices has been widely studied (Ashley, 2010; Ahmed, 2012; Verbeek and Groeneveld, 2012; Dobbin and Kalev, 2016; Baker, French and Ali, 2021; Buchter, 2021). Due to the overlap between the values which underpin both diversity initiatives and algorithmic bias mitigation, such as social equality, and in the absence of organisational literature which focuses on algorithmic bias mitigation, studies of diversity will serve as the basis of the discussion presented here.

Jenson *et al.*'s (2009) study on ethical codes of practice within business organisations, describes how these codes of practice form a complicated touchstone within an organisation. While these codes of practice are often generated to reduce moral ambiguity and provide clear direction when workers are faced with moral dilemmas, these are often decoded and understood in different ways across an organisation (Jensen, Sandström and Helin, 2009b).

This echoes Sara Ahmed's (2012) work on diversity initiatives, looking at how insider diversity activists do diversity work, and their use of bureaucratic instruments such as organisation strategy documents, audits, regulations, and committees (Ahmed, 2012). Within her extensive work interviewing diversity activists, she found mixed opinions as to the usefulness of bureaucratic instruments within their own organisation. Whereas some of the activists she interviewed used these instruments within their work and found them useful for holding organisations to some degree of accountability, others had concerns they became more of a box-ticking exercise (*ibid*). While organisational change in these circumstances often involves the production of procedures, organisational policy, and mission statements, these alone are not enough to allow an organisation to successfully transform their working culture (Ashley, 2010). Instead, organizational commitment to diversity policy has been shown to rely on insider activists to push forward and use the organisational tools available to them to advocate for change (Buchter, 2021).

The above demonstrates a complex r"lati'nship between ethics-based organisational change, regulation, policy production, and cultural working habits, operating on numerous organisational levels at once. In my own research, it was important to consider the character of the civil service organisational context within my methodological approach. This will be

discussed further in the following section, where I discuss the methodology utilised for this study.

6.3. Methodology

This study addressed the following research questions; what type of approaches would work to mitigate the effects of algorithmic bias within the DWP? And what are the challenges and opportunities to bias mitigation within the DWP?

To investigate these issues, I conducted a qualitative study in two parts. During the first part, I held seven online educational workshops at the DWP. In the second part of the study, I conducted follow up interviews with seven workshop attendees, to understand which parts of the workshop were most salient to their working lives, and how the methods to which they were introduced in the workshops might be integrated within their organisational context.

The first six workshops were designed as a multi-part educational series, the content of which can be seen below in Fig. B. The first four workshops were designed to introduce participants to different approaches to algorithmic bias mitigation, with the final two workshops providing attendees the space to think about how these concepts might be put into practice. After the original series of six workshops, one of the participants approached me about running an extra session for another group within the DWP, due to growing organisational interest in this area. This workshop (workshop seven) contained content from workshops one to four in an abbreviated format, and the algorithm prototyping exercise in workshop five. Workshop participants were invited to take part through an email describing the contents of the workshop series, and also by word of mouth, facilitated by key contacts at the DWP.

The participants targeted were from a range of positions across the department, with the idea of bringing together participants from data science, policy, and operational staff such as work coaches. This was due to my previous research in Chapter 5, which suggested the importance of practitioners from across an organisation or project developing a shared understanding of algorithmic bias. This decision was made between key contacts at the DWP and myself. My DWP contacts then contacted people in the organisation who they thought would be interested, either by emailing them individually, or emailing them through groups. These emails contained a small summary of the contents of the workshop. Participants who came to workshops 1-6 were primarily from the data science department, as well as one participant from policy, and one from central analysis. Participants were not asked which department they came from as part of the information and consent process, so I am only aware of my participants through the data collected during the workshops. Workshop 7 was attended by a much broader range of participants, including data scientists, cyber security, a member of DWP central diversity team, policy analysts, and Jobcentre work coaches. Workshop 7 was advertised differently, with a different contact taking on the role of contacting potential participants. This contact had previously attended some of the workshops in the original one to six series. Below in Fig. B is an estimate of how many participants attended each workshop. As some participants dropped out to attend other meetings, and participants sharing the link to the workshops with other colleagues, and it is therefore difficult to assess how many attended some sessions. Because of this, I reinforced at the beginning of the session, and

when people joined, that the session was being recorded and the data collected for an academic research project. Additionally, I stressed that if they did not consent to this, they would need to leave the workshop. Participants were also made aware that a recording of the workshop would be available on the DWP OneDrive for those who were not able to attend.

No.	Workshop session content	**Length**	**#participants**
1	**Workshop 1: Introductory session (WS1).** This session involved a talk on projects 1 and 2, as well as an introduction to project 3. 30mins Q&A and discussion.	1 hour	15 (Approx.)
2	**Workshop 2: Designing with values in mind (WS2).** This session involved activities based in value sensitive design methods, focused around stakeholder analysis and engagement.	1 hour	8 (Approx.)
3	**Workshop 3: Impact assessments for data-driven technologies (WS3).** This session allowed participants to explore new impact assessment standards being developed to mitigate the impact of bias in data-driven technologies.	1 hour	6 (Approx.)
4	**Workshop 4: People behind the datasets (WS4).** This session focused on the people behind the numbers, using creative story-telling exercises and ethical vignettes to explore issues of bias.	1 hour	7 (Approx.)
5	**Workshop 5: Algorithm prototyping session (WS5).** This session focused on prototyping an algorithm which effectively mitigates risks of bias.	2 hours	7
6	**Workshop 6: Framework prototyping session (WS6).** This session involved participants creating their own framework for mitigating bias when using algorithmic technologies at DWP.	2 hours	2
7	**Workshop 7: Compressed workshop series session (WS7).**	2 hours	18

Fig. B. (see Appendix C for a fuller draft of each workshop)

The workshops focused on how algorithmic bias develops, using high-profile cases as examples, a design choice informed by McNamara *et al.*'s (2018) study as discussed in Chapter 2, in addition to educating participants on interventions such as algorithmic impact assessments, value sensitive design, and critical thinking about data. In common with focus groups, workshops include interaction between participants, and allow opinions to be

revealed which might not otherwise have surfaced in a traditional one to one interview (Morgan, 1998).

In addition, the workshops provided a space in which practitioners of different types could come to a shared understanding of the issues they were discussing (Morgan, 1998). This was an important consideration while designing my methodological approach, as the findings of Chapter 5 (paper two) suggested that some of the difficulty in mitigating algorithmic bias came from a lack of shared understanding amongst practitioners. As noted in Chapter 3 above, using workshops as a research method created a space where participants could jointly identify and articulate language around 'fuzzy' issues, as well as develop a shared sense of understanding and communication (Ørngreen & Levinsen, 2017). Furthermore, it is argued that this type of space allows for tacit knowledge to become more visible, as participants are required to communicate with each other about their own workings and assumptions (Freytag and Young, 2017). This reasoning informed my decision to have the workshops open to a wide selection of DWP practitioners. Participants were invited from a wide range of backgrounds and knowledge domains within the DWP, and included data science practitioners, work coaches, diversity specialists, and research co-ordinators – the only criterion for attending was an interest in algorithmic technologies and discrimination.

The educational material for workshops two to four was developed during the analysis process of Chapter 5 (paper two), in which I identified algorithmic bias mitigation methods which would be suitable for inclusion in the workshops. These included Value Sensitive Design (VSD), algorithmic impact assessments, and critical thinking skills regarding data. I planned a workshop around each of these methods, by reviewing the literature in these areas in addition to identifying activities which had already been designed for those methods.

For the VSD workshop (Workshop 2), I included a direct and indirect stakeholder analysis exercise, as well as a value source analysis exercise, to provide an introduction to the VSD framework (Friedman, Hendry and Borning, 2017). This stakeholder analysis exercise involved participants being asked to read an ethical vignette focusing on a fictional algorithmic technology being developed by a fictional welfare department. The fictional algorithmic technology was being designed to match unemployed social security claimants to jobs for which they might be suitable. After reading the ethical vignette, participants were asked to identify the direct stakeholders (those directly involved with or impacted by the technology) and indirect stakeholders (those indirectly involved with or impacted by the technology). Participants were also asked to identify the values within the fictional algorithmic technology, in addition to imagining what the values might be for the stakeholders they identified.

For the algorithmic impact assessment workshop (Workshop 3), participants were asked to fill in the Canadian Government's Algorithmic Impact Assessment Tool online using information from an vignette focusing on a fictional debt repayment technology, which was based on the algorithm in the Robo-debt case study investigated by Park and Humphrey (2019; Park and Humphry, 2019; Treasury Board of Canada Secretariat, 2021). In the critical thinking workshop (Workshop 4), participants were given a public sector dataset to explore, and were asked to explore their own assumptions about what the dataset was about and who

it belonged to. Participants were also encouraged to consider how their assumptions related to social inequality.

The algorithm prototyping session (Workshop 5) involved participants rapid prototyping an ethical algorithm, a process which involved identifying the algorithm's scope, what data would be used, and what the business case was for this algorithm. The activities in this session repeated some of the exercises from sessions two to four, although in this session participants were asked to use these exercises to assist in prototyping their algorithm. In the framework prototyping session (Workshop 6), participants were asked to read through the CDEI's data ethics framework and discuss their thoughts on it (*Data Ethics Framework*, no date). The slides and activity Jamboards for the workshops can be found in Appendix A and B.

Workshops were conducted online using Microsoft Teams, as this was the DWP's preferred video platform. Regarding the format of the workshops, in the first workshop, I gave a talk providing an overview of algorithmic bias, in addition to presenting the findings from the first two papers of my thesis, followed by a 10-minute questions and answers segment. The format of workshops two to four was that each workshop began with a 10-minute presentation on the focus of the workshop, followed by learning activities facilitated by Google Jamboard. Depending on the number of attendees, participants were assigned breakout rooms to keep group numbers between three to five participants, with the aim of providing participants with groups small enough small that all participants were able to contribute towards the discussion. Workshops five and six followed a similar format, except they were two-hour sessions.

The data generated from these workshops came In two types. The first, with permission from the participants, was the video recordings of the workshops using Microsoft Teams' built-in video recording software. The second was the Jamboard exercises, which provided a visual record of participants' engagement with the workshop content.

Between one to three months after the workshops, depending on the participant, participants were invited to take part in follow up interviews. These were conducted using a semi-structured format. The interview guide for these interviews was developed based on analysis of the workshops. It included questions about which aspects of the workshops participants had found most relevant to their working practices. Participants were asked to recall specific examples of times during their working practices they had thought back to the workshop material, to ground answers in the specifics of participants' own lived experience (Mason, 2002).

These interviews were designed to understand which aspects of the workshops had been most salient to participants' everyday working lives, and what their thoughts on algorithmic bias and the workshops were more generally. Interviews were often with experts, some of whom had PhDs, and some of whom were in management roles. As noted in Chapter 3, due to the position which these participants had within the DWP, these interviews can be understood as expert interviews. Expert interviews are useful for understanding how the framing of particular problems might be influenced by expert opinions, and understanding the sense making which goes into creating those framings (Bogner, Littig and Menz, 2018).

After the workshops and interviews were conducted, transcripts were created using a combination of Teams' auto-generated transcripts function, and my review and correction of the transcriptions. Data analysis was undertaken using both deductive and inductive coding. Since I was using organisational theory as a theoretical framework to understand how organisational change occurs, some codes were predetermined using this framework (Braun, Clarke, and Hayfield, no date), such as 'opportunities' and 'barriers', as well as codes for 'future plans for mitigating algorithmic bias'. Additionally, I used deductive codes to assess the different levels of the organisation being referred to by my participants, such as 'ecological', 'organisational', and 'socio-psychological'. As part of this analysis, I went through the traditional process of familiarisation, coding, searching for themes, reviewing themes, and defining and naming themes, and then writing up my findings (Braun and Clarke, 2006).

Prior to collecting data, ethical considerations were discussed with my contacts at DWP and my supervisors. Ethical considerations included ensuring participants had contacts in case any of the material in the workshops upset them, data security, concerns regarding confidentiality, and ensuring participants were treated with respect and care throughout the data collection process. Ethical approval was secured from the University of Sheffield, and the application was approved prior to instigating the workshops (See Chapter 3 for fuller discission of methods and ethics).

6.4. Findings

The findings below are presented in four sections. The first, 'Perceptions of the diverse values of governments and publics', focuses on the tensions practitioners experienced when trying to balance the assumed values of a large diverse population in relation to each other. The second, 'Responsibility for embedded values', focuses on how practitioners felt uncertain about their responsibilities regarding algorithmic technologies. The third, 'laws and guidelines culture', focuses on how participants described a culture which emphasises following legal frameworks and guidelines, and expands on the findings on legal culture found in Chapter 4. The fourth, 'diversity', focuses on how participants felt about increasing team diversity as a potential way of mitigating algorithmic bias, and whether this may be another route towards algorithmic bias mitigation.

Perceptions of the diverse values of governments and publics

Using a sociotechnical framing of algorithmic bias, embedded values are understood to play an important part in the algorithmic bias mitigation process, and different methods have been proposed to assess how values are embedded and perpetuated by technologies (Friedman, Hendry and Borning, 2017; Eubanks, 2018; Selbst *et al.*, 2018; Lüthi, Matt and Myrach, 2021). This section explores how practitioners thought about values within the context of their work, in addition to discussing the conflict which arises due to the discrepancy between values held by the Department for Work and Pensions as an organisation and the values assumed to be held by the public.

In the workshop on Value Sensitive Design (Workshop 2), participants were asked to read an ethical vignette, which described a fictional algorithm being trialled by a fictional government welfare department. The fictional algorithm was being designed to provide jobseekers with recommendations as to the types of work they might be suitable for. After reading through this, the participants were asked to complete a stakeholder identification exercise, in which they were asked to identify which stakeholders would need to be considered when designing this algorithm. In response to this exercise, some participants discussed the importance of the organisations' values in relation to 'the taxpayer'. When participants reflected on what would or would not be fair in response to hypothetical examples of algorithmic projects, the taxpayer often featured as the central discussion point. Frequently, this was expressed in a similar vein to the participant below:

> "[Y]ou want to be as efficient as you can with taxpayers' money, because these are hard-working individuals who give money to provide a service and you want to make sure that you're respectful of that." (P1, Data Scientist)

Another participant expanded on this, explaining that they understood the taxpayer to be the one who paid for the service provided by the DWP, and also that deemed it necessary to balance the taxpayers' against those of claimants, since the department, in their understanding, was spending the taxpayer's money:

> "So, in a sense, every pound that we pay out to people in benefits is a pound that's taken from other people in taxes – very simplistically. So, you need to be fair to both sides…. [..] how much weight you put on one set of values, being fair to the people who are direct customers, versus how much you put in another set of values, which is how fair we are to the people who pay for these services." (P4, Analyst)

As these participants reflected on this, they appeared to treat the claimant and the taxpayer as two discrete categories. Moreover, the taxpayer in these conversations was often conceptualised as an income tax paying citizen, rather than corporations who pay tax, or citizens who pay VAT. In doing so, the fact that someone's claimant or income tax-paying status often changes across their lifetime was overlooked. Those who are currently paying income tax may previously have claimed benefits, and those who are currently claiming benefits may previously have paid income tax. Furthermore, income tax is not the UK's only source of tax revenue. In 2022/23, VAT, a tax which is paid by all citizens regardless of their working status, was the third largest revenue source after income tax and national insurance contributions – generating £160bn towards the UK's annual revenue (Keep, 2023). Additionally, my participants' conceptualisation of taxpayers and claimants as two distinct groups overlooked how the social security net works across the course of a citizen's life, funded by National Insurance, paid by all citizens in case of illness, circumstance, and old age. It is worth noting that not all participants mentioned the importance of balancing claimant and taxpayer needs, however most participants did at some point refer to the importance of the taxpayer when discussing values within the context of their work.

However, while participants perceived cost effectiveness to be a value the taxpayer held, they also imagined there was diversity in the opinions of taxpayers regarding how the DWP should

be run. Furthermore, participants admitted it was difficult to assess what the values in groups as large and heterogeneous as taxpayers and claimants might be. One participant, when discussing his perception of the values of the public, said:

> "[D]ifferent stakeholders have different requirements, the taxpayer probably wants you to accept the first… or some taxpayers would want you to accept the first job offered to you, like any job's a job, sort of mentality. Whereas […] nicer citizens might be like, you know, we want you to take a job that you're happy in." (P1, Data Scientist)

During the follow-up interviews, the perceived taxpayers' opinion was often referred to when discussing how the department is run. Participants stated that ministers and politicians were the ones who moved the department along, as well as being the ones responsible for ensuring the department met the taxpayers' needs. Respect for the taxpayer, and the democratic structures which gave ministers their responsibility towards the taxpayer, was clearly valued amongst many of my participants.

Costanza-Shock's (2020) discussion of value sensitive design, states that when data scientists favour their own perspectives they subjugate those already marginalised (Costanza-Chock, 2018; Green, 2021). They also assert that technology must be designed around marginalised groups' needs in order to mitigate the effects of algorithmic bias. Their discussion then uses Freidman and Nissenbaum's (1996) taxonomy from *Bias in Computer Systems* to describe how already present institutional values become hard coded within algorithmic systems, known as preexisting biases. Furthermore, Costanza-Shock argues that these preexisting biases must be addressed to mitigate bias in algorithmic technologies (Friedman and Nissenbaum, 1996; Costanza-Chock, 2020). In the approach advocated by Costanza-Chock, the consideration of marginalised groups, and analysis of the wider dynamics which lead to these groups marginalisation, is at the heart of social justice based approaches to moral dilemmas (Dencik, Hintz, and Cable, 2016).

However, my findings suggest there are limitations in the extent to which these approaches can be applied in a DWP context, due to the ways in which the organisation is embedded within its broader social and political structures – that is, due to the importance of the ecological organisational level. Or, to put it another way, the DWP is limited in how it can utilise social justice-based approaches due to its UK civil service context. Within this context, civil service departments operate in a manner dictated by the political structure of the UK, which is based on a form of representative democracy, where government spending and policy priorities are decided by ministers in part to satisfy campaign promises and win subsequent political campaigns (Guinaudeau and Guinaudeau, 2022). Moreover, political parties must secure a majority share of the votes during an election to form a government – meaning that to a degree, they rely on popularity within the majority. This suggests practitioners may face tensions in their roles as civil servants, in trying to centre technologies around marginalised groups while at the same time serving a government which relies on popular appeal – a point I return to in the next paragraph.

Relatedly, participants discussed how the values present in an algorithm might result from policy, or the influence of political lobbying. For example, in a follow up interview with a policy

analyst, this participant described how some lobbying groups held more influence than other lobbying groups. When discussing The Bedroom Tax[7], this participant described how the lobbying groups representing the armed forces held more sway than those representing disabled people, and thus were more successful in gaining an exemption to the new policy. Expanding on this observation, this participant noted that "certain groups have more political sway than others. So, [they're] more likely to get things written into your algorithm too" (P7). Again, this suggests the significance of the political system within which government civil service departments are embedded at the ecological level when attempting to mitigate algorithmic bias. Those most at risk of algorithmic bias within a UK welfare context come from already marginalised groups within the UK political system, a system which directly influences how civil service departments operate. This suggests there is work to be conducted to scope out what work civil service data practitioners have the agency to do if they want to centre marginalised groups in the development of algorithmic technologies within the current political structure.

Allied to this, participants saw the DWP as being under pressure from the media within the context of their work. DWP is often mentioned in news media, with articles focusing on topics such as benefit-related suicides and food bank use, reports of benefit fraud, and sensationalised content about criminals receiving benefits (Alibhai, 2019; *BBC News*, 2021; Butler and editor, 2022; Heffer, 2022). Although the news media in the UK is diverse and varied, and not always critical of the DWP, participants nonetheless felt that the media could at any moment criticise any DWP policy or (and sometimes falsely, according to my participants, with coverage of projects which had not even been planned or investigated). These news articles impacted department morale, with one data scientist participant mentioning he was anxious about the possibility of seeing new articles describing how another claimant had committed suicide while waiting for benefits. Moreover, participants felt portrayal of the department could fluctuate between either wastefully left wing or cruelly authoritarian. My participants perceived the department as being influenced by the types of stories written about them, at an ecological organisational level. When reflecting on the department's future plans for algorithmic technology, one participant said:

> "I think, any government department, it's pretty risk-averse on this sort of thing. So, like, you know, they'd rather not do something than risk something going wrong because you know everything you do is open to, like, media scrutiny and stuff." (P1)

According to my participants, members of the public who are "nicer people," as one participant put it, will be more upset by news articles about claimants' mental health, whereas those who are "spendthrifts" might be angered by articles saying claimants are being treated leniently (P1). The concept of 'nicer people' is problematic in this context, and research on the perceptions of citizens regarding public sector use of data suggests citizens have a diverse range of perceptions, as previously discussed in Chapter 2 (Ditchfield *et al.*,

[7] The bedroom tax is a colloquial description of a change in in public housing policy where tenants in public housing receiving payments towards their rent would receive a payment reduction if their house was 'underoccupied,' meaning having more bedrooms than there were tenants.

2022). Moreover, as civil servants, some participants felt the need to respect the diverse values they perceived the public to hold, including both those whose primary concern was that the department did not overspend, and those who were concerned about harsh policies causing claimants to have mental health difficulties. Participants said this was important even when they disagreed with the values they perceived the public as having, and that it was necessary to create a sense of distance between their personal values and those of the department.

Because of these diverse factors, participants saw substantial difficulties in assessing what would constitute the right or wrong values to prioritise within the context of their work. During one of the follow up interviews, one of the data scientists reflected on the difficulty of knowing what was fair in the context of the organisation's work:

> "[O]nce you get into murkier waters… we're talking about people's perception of fairness, right? And again, everyone is different, and we'll have different backgrounds, different norms and values, and there we go…. We've got many definitions of fairness, so how do you then get an objective view of what is fair?" (P2, Data Scientist)

Jenson *et al.* (2009) argue that guidelines and codes of practice can only take someone so far in ethical conduct. Moreover, they identify 'the moral dilemma' as something which only occurs when it is unclear what the right course of action might be (Jensen, Sandström and Helin, 2009b). The sense of uncertainty about what is 'fair' felt by my participants was exacerbated by many of the DWP's 'products' serving large populations – with different needs, imagined values, and expectations and beliefs about how a society, and its welfare service, should function. Using Jensen's understanding of the place of uncertainty in ethical behaviour, I suggest that the organisational culture of legal and policy compliance found within my research (and also discussed in Chapter 4) may encourage an expectation of formalizable processes regarding moral dilemmas. In Jensen *et al.*'s (2009) view, for someone to engage in moral action, they must practice self-reflection and have a sense of personal responsibility for the consequences of their own actions. Jensen *et al.* (2009) also argued that this is not encouraged by organisational cultures which favour strict adherence to codes of conduct (Jensen *et al.*, 2009; Dillard and Yuthas, 2002).

Responsibility for embedded values

As can be seen in the previous section, when considering the values embedded in algorithmic technology, it is difficult for civil service practitioners to adopt social justice framings due to the political structures the department is embedded within, in addition to the diverse values which are perceived to be held by the public. This section focuses on what my participants felt their personal responsibilities were regarding the values embedded within the algorithmic technologies they worked on.

Due to the self-selecting nature of my participants in this study, it is perhaps unsurprising that nearly all were aware of the ethical and political tensions which existed within their own work and that of the department more generally. As discussed previously, the ecological

organisational level creates difficulties for practitioners to assess how to approach moral issues within the context of their work, because civil service practitioners are required to enact the policies of the current government. In conjunction with this, participants described the importance of team support when broaching moral issues within their work. Talking about his experiences when being asked to work on a project he felt was of moral concern, P6, a data scientist, said:

> "I've been quite lucky in that I've always worked in teams that have had a good leader, basically able to go back and say like, no, this has got to be pushed back up […] But it's very much down to I think, the individual... if you get someone who's good at it [who's] like, "we're just not doing it", but like, [who's] quite tactful and basically just like chaperone in that way." (P6, Data Scientist)

This suggests the importance of relationships at the socio-psychological level of the organisation, and the significance of individuals who can sensitively approach topics of moral concern.

However, participants did not see all uses of algorithmic technologies as being of moral concern. Some data scientists expressed concern that the public imagination had been narrowly fixed on the most harmful examples; on the "truly horrifying car crash stories" (WS7), such as the example of the US criminal justice algorithm identified in the ProPublica investigation and described above (Angwin *et al.*, 2016) While all participants sought to prevent similar issues occurring within the context of their work, some described their work as often far more mundane. Examples of these mundane situations included algorithms to sort through emails or filter survey responses.

Some of these participants were concerned that by only focusing on the most harmful examples, there was the potential to create a misunderstanding that all algorithmic or data driven systems had the potential for harmful or biased outcomes. Furthermore, some participants were concerned that the focus on harmful examples might contribute towards practitioners working on algorithmic technologies having more responsibility for the decisions made by the algorithms they developed than human-decision makers for similar decisions.

Participants from various sections of the organisation were aware of how cases of human-based discrimination and bias might develop within their departments. During the algorithm prototyping session (Workshop 6), a participant said that they would feel uncomfortable using certain DWP data sources for their hypothetical algorithm. This was because these data sources contained human made health assessment decisions, the decisions that could be overturned at a later date due to additional evidence. Conversely, a participant who worked on the diversity team mentioned that she felt cases of human discrimination still did not receive enough attention within the department. This was also mentioned by a participant who was a work coach, who commented that he thought the Universal Credit system – a system designed by human decision makers – contained biases against migrants and part time workers. Although these decisions were made by humans, my participants did not think this made these decisions 'fairer.'

These reflections point to three complications in the comparison between algorithmic and human decision making. One, algorithmic technologies are often trained on human decision making. This point has been well documented by academics regarding supervised machine learning (Balayn and Gürses, 2021; Jaton, 2021; Pessach and Shmueli, 2023). When an algorithm is given biased data to learn what is a 'correct' or 'incorrect' decision, it will learn the biases of the label assigning system. However, a less discussed point here is that if the algorithmic technology's outputs are based on human decision makers' judgements, then removing the algorithmic technology will not prevent systematic bias. Specifically, if little attention is paid to systematic bias across human made decisions in an organisational context, and little progress is made towards removing these biases in human decision making, then the net result might be said to be quite similar to the effect of algorithmic bias – marginalised groups are still systematically discriminated against. I suggest this indicates the importance of approaching human bias within decision making processes alongside algorithmic bias mitigation methods. To be clear, this would mean addressing not just the human biases of data scientists, but also those of human decision makers who make the types of decisions which might be automated.

Direct comparisons between human and algorithmic decision-making processes have been scarce. Dressel and Farid (2021) investigated algorithmic versus human decision making in the context of recidivism risk assessment, risk assessments in the criminal justice system which look at a defendant's likelihood of reoffending. The researchers compared the judgements of naïve participants, by which they mean those who were not legally trained on the judgements produced by recidivism risk assessment algorithms (Dressel and Farid, 2021). The researchers found that naïve assessors performed similarly to the algorithmic risk assessment tool. However, perhaps more interestingly, the authors also mention that they approached judges in the criminal justice system to participate in a similar study. The similar study would involve the risk assessment algorithm's judgements being compared to participant judges' assessments. However, their request was denied; according to one judge's clerk, this was because the judge was concerned they would perform no better than, or worse, than the naïve assessors in their study (*ibid*).

There have been calls for data scientists to be more realistic in their assessment of the abilities of algorithmic technologies (Green, 2020). This is because some data scientists, and some organisational managers, over-estimate the accuracy of insights which are possible, positioning these technologies as more objective and intelligent decision-makers than human decision-makers. Green (2020) states that employers and data scientists often over-estimate what algorithmic technologies can do. There is an adjacent need to ensure that algorithms are not judged by a higher standard than the human decision makers within the decision-making context. I suggest there is a need to view algorithmic technologies alongside human decision making, and consider these two issues in tandem, and not as two opposing issues. Or, to put it another way, it is important that algorithmic decision making is not judged against a non-discriminatory ideal, rather than the options realistically available to organisations. This suggests the need for careful consideration as to how algorithms are judged, and for further research into how they compared to human decision makers in real life contexts. This issue

of comparison to human decision-making processes will be returned to again in the following section.

Laws and guidelines culture

Outside of value-based approaches, algorithmic bias mitigation can be approached through a legal perspective. This theme focuses on how participants engaged with and understood laws and guidelines within the context of their work, and notably, how some participants felt certain laws were unclear regarding their professional responsibilities.

As Orr and Davis (2020) found in their empirical work on attributing ethical responsibility with AI practitioners, practitioners relied primarily on legal structures to decide what is fair within their working practices. This section suggests that legal frameworks not only provide a minimum standard to adhere to, but also impact the actions practitioners and organisations may see as feasible when looking to mitigate algorithmic bias. My participants were incredibly positive about algorithmic impact assessments (AIAs) and saw them as similar to other assessments they use within their work. For example, they felt that AIAs were similar to DPIAs (Data Protection Impact Assessments)[8]. The department is already legally accountable for producing DPIA documents. Documentation such as DPIAs and AIAs have been criticised as potentially being box ticking exercises, as they are detached from the harms they are meant to assess (Metcalf, *et al.*, 2021; Yam and Skorburg, 2021). However, participants described mechanisms such as these as important in providing them with a skeletal framework with which to consider issues during the development process. For example, in workshop 6, in which participants reflected on frameworks for mitigating algorithmic bias, one participant stated he felt "the idea of the framework is basically to set out seemingly obvious things, just to put it on paper, so you put most rules and laws down just so, you know, it's been considered and it's a general sort of [don't] take anything for granted" (WS6), rather than to prescribe any particular set of action.

Participants were aware of regulatory requirements for a 'human in the loop' for projects deemed as being 'high risk,' in the GDPR. The 'human in the loop' requirement requires data controllers to ensure decisions are not solely made on the basis of an algorithm – for there to be a 'human in the loop.' However, some participants questioned how much this really improved decision making. Some participants expressed that, in a DWP context, humans might assume algorithmically generated decisions were correct and defer to them, as was found during Eubanks' study of US public sector workers (Eubanks, 2018). Participants shared various experiences with human and machine generated recommendations, including where people had ignored the machine generated recommendation altogether, or only followed the recommendation so they would not get in trouble for ignoring them. Similarly, during conversations, some participants challenged the notation that following recommendations without critically scrutinizing them was only of concern within the remit of algorithmic decisions. Instead, participants saw this as an issue with recommendations more generally,

[8] An assessment process designed to identify and minimise the risks associated with data use on a project and required by the GDPR if a project is likely to be of high risk to individuals.

including recommendations provided by other human beings. One participant discussed this within the context of human produced assessment scores, which were produced for the DWP by a third-party organisation independent of the department for the purpose of assessing whether Personal Independence Payment (PIP)[9] applicants met the criteria for the benefit. He spoke about his concerns regarding whether there was any meaningful intervention from case managers once they received the assessment scores:

> "the thing is that there is a related issue in health assessments, which is that technically, all decisions on whether to award people benefits are made by DWP decision makers, not by the assessment service, but as far as I know, the proportion of cases where decision makers do something other than accept the recommendation is vanishingly small." (WS5)

Returning to the issue of legal frameworks, participants expressed a lack of certainty regarding how to apply anti-discrimination laws to their own work. One participant, an analyst, recalled a trial the department had conducted which aimed to offer more support to people who had not been able to obtain certain educational qualifications. This trial had been cancelled due to potential unequitable treatment across protected characteristics, as analysts found there was a large overlap with variables relating to educational qualifications and race. The participant recalled that it had been asked whether the intervention would be indirectly discriminating against people based on race, using education as a proxy. Although this intervention would be providing support for these particular groups, there were still concerns this would be discriminatory. In response to this, one participant suggested The Equality Act (2010) might provide a useful definition of how to treat different groups fairly. However, after some deliberation, both participants agreed that "that's fairly vague as well" (WS6). While there have been calls for greater regulatory protections in relation to algorithms, my participants were more concerned with the lack of clarity they perceived within anti-discrimination law, and how this applied to using statistical methods. As participants were strongly influenced by legal frameworks such as The Equality Act (2010), this may suggest the need for civil service departments to provide their employees with greater clarity around discrimination law.

Additionally, while criticisms of procedural and statistical parity-based methods have been influential within algorithmic bias mitigation communities (Green and Hu, 2018; Hoffmann, 2019), it is important to consider these methods in relation to how practitioners perceive legal limitations. While the law makes exceptions for unequal treatment if it can be proven this is to achieve a "legitimate aim," it may be difficult for those who are not legal experts to understand where these exceptions lie and to apply the law with confidence. To further complicate matters, within UK discrimination law, what constitutes a "legitimate aim" is usually uncovered during case law. An example of a legitimate aim might be a company hiring someone on the basis of them speaking Hindi due to the company's work with India overseas. While it could be argued this is indirect discrimination against those without Indian heritage,

[9] PIP is a UK social security benefit targeted at people with disabilities or health conditions which significantly impact their day-to-day life.

the employer could argue this criterion was a proportionate means to a "legitimate aim". In this case, the legitimate aim would be a "genuine business need" (Acas, 2023). In this context, risk averse organisations may be reluctant to act if there is any uncertainty as to the legality of their actions.

Diversity

As discussed earlier, diversity initiatives are where an organisation aims to either recruit more employees with certain characteristics, or promote employees with certain characteristics who are underrepresented in the organisation at a senior level. Although I did not discuss diversity initiatives as a method of algorithmic bias mitigation, and did not ask about them during interviews, participants often discussed diversity initiatives as a pathway for algorithmic bias mitigation of their own volition. This section focuses on how my participants saw diversity within algorithmic bias mitigation, specifically around a) how more diverse teams might be able to offer greater insight into the types of discrimination an algorithm might produce, and b) how difficult it is to increase diversity within the department.

Expanding the diversity of data science teams has been posited as one way of mitigating algorithmic bias (Kuhlman, Jackson and Chunara, 2020). It is argued that the lack of diversity found in data science teams makes it challenging for these teams to interpret data in a way which does not perpetuate social biases (*ibid*). By having more diverse teams, Kuhlman, Jackson and Chunara argue, the gap between data scientists' perceptions of marginalised communities and the communities themselves could be decreased (*ibid*). Participants felt that having more diverse teams engaged on their projects would help them spot biases earlier on in the project, and would help make projects more inclusive. For example, talking about team diversity, one white male data scientist said:

> "Certainly, in the team that I'm in at the minute, it's not very diverse at all…. As a group, it's really hard to like have an idea of your unconscious bias, but unless you have that diversity, you don't know what you're missing. You don't know what you don't know, you know? So, it's like unknown unknowns, you have an idea of it, but you don't get necessarily what other people's lived experiences are." (P6, Data Scientist)

However, despite hope that more diverse teams might fix the issue of algorithmic bias, this is not without its own challenges. One participant, a black queer work coach who helped claimants on the operations side of the business, said he wanted to "go up the grades" to improve things for minorities from the inside (p3). He argued that it was incredibly important that the DWP had more diverse voices at higher levels, to ensure fairer and more just decision-making processes within the organisation. However, during our conversation he also expressed uncertainty as to whether he would be able to make any difference at all, saying that:

> "I even feel if I even get up near there, I'm still not gonna be able to change much because you kind of become part of the fabric of where you're working" (P3, Work Coach)

As Costanza-Chock argues, having diverse designers of systems still often leads to projects which have the majority or dominant culture in mind (Hamraie, 2013; Costanza-Chock, 2020). A potential explanation for this is that marginalised designers design for the majority dominant culture because they are continually exposed to it, and so struggle to envision what technologies designed for marginalised users would look like (*ibid*).

Furthermore, it has been noted that for minority groups to be allowed access to particular spaces, they may have to adopt certain stances, viewpoints and behaviours in order to be considered welcome and viewed as the right "fit" for the organisation (Ashley, 2010). These expectations may exclude those who are most likely to bring a critical eye to a project, and instead favour those best able to "play the game" (Carbado and Gulati, 2000; Ashley, 2010). One participant, who worked on the diversity team, said she often had to explain to diverse employees that the organisation cannot change as fast as they would like, and they would need to be patient and play the "longer game" in convincing others to make the necessary steps to increase diversity in the organisation (P5). By this the participant meant she thought marginalised employees often wanted change to occur faster than was possible within the organisation, and by pushing for things too soon, there was a risk of making the situation worse. This participant, who works in the diversity team, said:

> "So, I know sort of from when I presented the long term view of trying to get sort of movement in race, which is quite difficult as a white woman […] when you're in the minority, and you're kind of going "I'm trying to help you here, but we need to stop talking about race quite as much, because otherwise we're making things worse" (P5)

In Workshop 7, a data scientist commented that those in attendance were "closer to the Silicon Valley types which invented [these] algorithms" (WS7). By this, he meant many at the workshop were either white, affluent, well educated, or men. Similarly, some participants expressed they couldn't imagine this overrepresentation of particular demographic groups would be solved "any time soon" (WS1). One participant, who worked within the DWP's diversity team, explained that movement in this area had been slow in the organisation. Not only did the department have difficulty when it came to hiring people with certain protected characteristics, such as black people and people with certain disabilities, but there were also difficulties when it came to promotion and retention. Due to The Equality Act (2010), hiring or promoting on the basis of having a certain protected characteristic – even if a marginalised one. The two exceptions are if it is to meet a "legitimate aim" (as described above), or in the case of two applicants being equally well matched to a vacant post, the organisation can choose to appoint an applicant from an under-represented background because that group is under represented (The Equality Act, 2010). Organizations can instead engage in what is known as positive action, which is encouraging certain groups to apply for jobs, although all applicants still must meet the job application criteria.

These insights suggest participants felt that greater diversity could improve the organisations' approach to algorithmic bias mitigation, although they did not feel that this was an approach which could be actioned within the short term. However, this approach may potentially divest non-marginalised practitioners from the responsibility of reflecting on their own social biases. Instead, it makes future marginalised practitioners responsible for closing the interpretation gap between data scientists' perceptions of marginalised communities and the communities themselves. Moreover, this not only places this responsibility on individuals who do not currently work in the relevant role, but on individuals who may already be struggling to gain work in that role due to social biases in the workplace. In other words, it makes future marginalised practitioners responsible for algorithmic bias mitigation, when this should be the responsibility of all department employees, marginalised or not.

6.5. Discussion and conclusion

In this paper, I have investigated how the DWP might implement algorithmic bias mitigation techniques such as AIAs, value sensitive design, and critical thinking skills regarding data. In doing so, this paper aimed to consider both the opportunities and barriers present in the organisational environment, and the implications this may have for public sector bias mitigation practice more generally. This study suggests that while algorithmic bias mitigation techniques such as algorithmic impact assessments and value sensitive design seem highly suitable for organisations such as the DWP, the implementation and potential impact of these techniques is difficult in practice. The findings from this paper assist in understanding algorithmic bias mitigation methods in the civil sector in three ways, by identifying two challenges and one opportunity within the civil service context. The first challenge is that it is difficult for practitioners to align technologies to the social justice values that underline socio-technical mitigation approaches when servicing a large heterogenous public. The second challenge is that practitioners' perception of a lack of clarity in anti-discrimination legislation can lead to uncertainty as to how conflicting needs within the population should be addressed. The opportunity presented is that participants were enthusiastic for more diverse teams as a means to uncover unknown biases within their interpretation of their data and data analysis processes. However, adopting this strategy has its own difficulties. I expand on these below.

Regarding the first challenge, the civil service context provides particular challenges in adopting algorithmic bias mitigation methods. This is due to the tension between social justice advocates' expectations of best practice (Costanza-Chock, 2018; Green, 2021), and the feasibility of this within a civil service context. Costanza-Chock's (2020) discussion of value sensitive design, argues that for practitioners to mitigate the risks of algorithmic bias the values of marginalised groups must be centred in the design process. My findings suggest that civil servants may struggle to utilise this approach for two reasons. The first rests on the challenge of designing technologies for a large and heterogenous population, who are perceived by practitioners as having differing viewpoints and cultural values. My participants expressed these concerns by describing how they believed that citizens hold conflicting values

by using taxpayer discourses – that is, narratives around the place of the taxpayer with regard to public sector services, and narratives around the perceived expectations of taxpayers. The second reason relates to how biases become embedded within organisational practice through the ecological organisational level, specifically through the organisation's relationship to the wider political structures within which it is embedded. Or, to put it another way, practitioners in a civil service context may struggle to centre marginalised groups in the development of their technologies due to government policy or influence from political lobbying groups.

Green (2021) proposes that there is a need for data scientists to view themselves as political actors and to reflect upon the political orientation of their work. In his argument, he provides an example in which 'tech workers' have protested against their organisations working with the US military and Immigration and Customs Enforcement (ICE), in addition to refusing to work on these projects in some circumstances (*ibid*). However, refusing to work on a project is more challenging within a welfare context, where refusal to work on projects in the department may lead to the public being underserviced in a critical public sector service. Furthermore, while civil servants at varying levels of the DWP do have some agency in their work, the stipulation that civil servants are publicly apolitical may limit the routes available for practitioners to protest against unjust projects in this context. Specifically, it prevents practitioners from engaging in any form of public protest against government policies which may increase the risks of algorithmic bias. This suggests that civil service data scientists may need to scope out the possibilities for potential action within their own organisational contexts, in order to identify what routes are available to them to effectively advocate against political influences which may lead to algorithmic bias. Specifically, this could mean addressing the following questions: what does centring marginalised groups look like in a civil service context? How might this be at odds with the structures of government it is embedded within and reliant upon? What other limitations might practitioners face in this space? In part, algorithmic bias develops due to the hard-coding of institutional values (Friedman and Nissenbaum, 1996; Costanza-Chock, 2020), and this raises the question of how far civil service practitioners can go in algorithmic bias mitigation without the radical transformation of the institutions they work within?

I now move on to the second challenge, that practitioners struggle to navigate guidelines and regulatory frameworks due to their perceived lack of clarity in these frameworks. My findings suggest that some practitioners find the definitions of discrimination provided in The Equality Act (2010) are not clear, which may lead to practitioners feeling uncertain of their options when attempting to mitigate algorithmic bias. Specifically, participants were concerned that The Equality Act (2010) prevented unequal treatment of people on the basis of protected characteristics. In Wachter *et al.*'s (2017) paper on automating decisions and the GDPR, they state that positive discrimination is legal in the UK and EU context if it is proportional and required to meet a 'legitimate aim' (Wachter, Mittelstadt and Russell, 2017). However, my findings suggest it is difficult for practitioners to judge what constitutes proportional or a legitimate aim. This raises questions surrounding how non-legal practitioners can be confident of whether their design choices do indeed constitute working towards a legitimate

aim and being proportional in achieving that aim. While critiques around the concepts of statistical and procedural fairness are rightfully important (Green and Hu, 2018; Hoffmann, 2019), these findings suggest that practitioners may face difficulties moving beyond these types of approach. This is due to practitioners' perceptions that regulatory frameworks require treatment to be equal, and so their work requires, at least in part, attention to be paid to statistical and procedural conceptualisations of fairness.

The findings presented here suggest it is important to recognise that practitioners' perceptions of these laws will shape their working practices and interpretation of legal frameworks. Social justice-based approaches to algorithmic bias mitigation often include a degree of positive discrimination to provide marginalised groups what they are 'owed' by the social contract, in addition to suggesting mechanisms for equality of opportunity. This suggests that for practitioners to become more comfortable operationalising social justice values, further clarity around regulations is necessary. This should be provided at an organisational level, so that practitioners are provided with examples relevant to their working practices.

Turning to the opportunity presented within my findings, participants perceived that the organisation's algorithmic bias mitigation efforts would benefit from having a more diverse workforce, both to assist in identifying harmful biases as well as to help implement algorithmic bias mitigation methods. However, while changes to organisational diversity practices may provide some assistance in mitigating algorithmic bias, it is important to remember this can be but one part of an approach. It is still important that all practitioners become more aware of how discrimination occurs. Furthermore, while organisations may want a more diverse workforce for the proposed benefits, such as assisting with issues such as discrimination and the organisations' diversity portfolio, this may be a heavy responsibility for individuals to take on. Specifically, for practitioners to assist in this way it requires them to use their lived experience of discrimination within the context of their working life, which marginalised individuals may not feel comfortable discussing with their non-marginalised colleagues. Additionally, for more diverse teams to be able to produce fairer outcomes, changes to organisational diversity policy may be needed to ensure workers feel confident and supported enough to make suggestions which go against the grain of current company culture.

This paper contributes towards answering my thesis' overarching research aim: to investigate how DWP practitioners might mitigate the impacts of algorithmic bias? The findings presented in this study suggest that while practitioners saw the value in adopting some algorithmic bias mitigation methods, in particular AIAs, much of the work necessary to work towards algorithmic bias mitigation will need to occur through the restructuring of relationships between the organisation, practitioners, and the wider ecological organisational level in which they are embedded. Additionally, while practitioners were eager to approach issues of algorithmic bias, they required assistance from other practitioners at the socio psychological organisation level to resist suggestions to do work they were not morally comparable with. These findings will be further examined in Chapter 7 (Conclusion).

Chapter 7: Conclusion

The motivation for my thesis was the rising concern surrounding the power that data and algorithms have to discriminate against minoritised groups, and how the public sector might seek to mitigate the impact of this problem. My three papers had the overarching research aim of exploring the challenges present in trying to mitigate algorithmic bias within a UK government department, the Department for Work and Pensions (DWP). A further aim was to investigate the way in which algorithmic bias mitigation practices could be incorporated into DWP practice. In this chapter, I provide a summary of each of my research papers and their empirical contributions. I then discuss the overarching contributions which my thesis makes to the emerging field of critical algorithm studies, and to fields interested in the influence of algorithmic technologies on society and culture, such as information studies, sociology, and communication studies. After this, I discuss the recommendations for practice which emerged from my research. Finally, I discuss the limitations of my research and point towards opportunities for further research regarding algorithmic bias mitigation in the public sector.

7.1. Summary of empirical contributions

In Chapter 4, *Investigating the role of current DWP working practices in mitigating algorithmic bias*, I investigated the working practices of DWP practitioners regarding algorithmic bias, and discussed the limitations of these practices. To do this, I focused on the following research questions; RQ1a: what algorithmic bias working practices are currently practiced by data science practitioners? and RQ1b: What are the limitations of these practices? To address these questions, I conducted eight interviews with DWP practitioners about two recent data projects on which my participants had worked; the Digital Trialling Framework and the Digital Plus Trial. These projects sought to develop a real time randomised controlled trial system which assessed claimants for a new DWP service, which was being developed to allow the department to oversee randomised controlled trials more efficiently. This study identified two key empirical findings. Firstly, my participants strongly relied on legal frameworks due to their position as civil servants, yet the legal frameworks they were required to adhere to did not facilitate accountability to the population they served. Furthermore, participants felt a strong sense of responsibility to citizens, but they perceived they were accountable to the state. Secondly, my participants' working practices in relation to bias checking were limited by previous research conducted by DWP, and by influences from the department's organisational culture.

Additionally, this paper contributed towards my overarching research aim; to investigate how DWP practitioners might mitigate the risk of algorithmic bias. The empirical findings presented in this study suggested the civil service context has its own challenges regarding algorithmic bias mitigation. Specifically, the organisational culture may place a different set of responsibilities on civil servants in relation to rules and guidance than those experienced in the private sector, and practitioners related this to their responsibilities as civil servants. This was then further explored in Chapters 5 and 6.

In Chapter 5, *Lessons in mitigating bias from the field: Exploring good practice and moral challenges on the Aurora AI project*, I investigated how practitioners on the purportedly progressive Aurora AI project by the Finnish Ministry of Finance were attempting to mitigate algorithmic bias. The paper presented in this chapter focused on the following research questions; RQ2a: What might 'good practice' on an algorithmic project look like? and RQ2b: What challenges does good practice on an 'ethical AI' project face in practice? To investigate this issue, I collected qualitative data through semi-structured interviews and utilised document analysis to understand how stakeholders on the Aurora AI project were responding to the challenges posed by algorithmic bias. Additionally, I interviewed AI Ethics experts, predominantly from algorithmic justice organisations, about the Aurora AI team's proposed algorithmic bias mitigation plans.

In the paper, I identified two key empirical findings. First, even in this purportedly progressive project, there was a lot of disagreement about what constituted good practice in mitigating algorithmic bias and the types of solutions that might be practically implementable. These differences in understanding, combined with systemic issues such as funding and organisational working practices, meant that more socially focused participants felt moral concerns were sidelined. Additionally, AI Expert participants perceived good practice regarding algorithmic bias differently. Participants from the Data Justice Lab placed a stronger emphasis on consideration of the wider ecosystem of inequality, and on approaches which de-centred technology (that is, focusing on how people are discriminated against outside of technological contexts), than the Ada Lovelace Institute participants. Moreover, the Ada Lovelace participants were more interested than participants from the Data Justice Lab in the use of AIAs as an instrument in mitigating algorithmic bias. The second key finding was that project management styles which focus on technological pursuits may not allow enough time to focus on how to mitigate the impact of biases. The project utilised an ethics committee as its primary instrument for mitigating algorithmic bias; however, members of the ethics committee were concerned that moral issues were sidelined in favour of prevailing value-based assumptions. These findings move beyond existing understandings of algorithmic bias mitigation practices which focus on either individual constraints or macro-level analysis, to address the importance of contextual organisational and structural constraints in public sector algorithmic bias mitigation.

Additionally, the research I undertook for this paper suggested three socio-technical algorithmic bias mitigation methods; Value Sensitive Design (VSD), algorithmic impact assessments (AIAs), and thinking more critically about how algorithmic technologies are used and developed; specifically, considering the wider context in which they are embedded. These socio-technical algorithmic bias mitigation methods were taken forward in the design of Chapter 6, where I investigated how practitioners perceived these might be incorporated within a DWP context.

In Chapter 6, *The influence of DWP organisational culture on the adoption of algorithmic bias mitigation practices and implications for practice*, I discussed the results of a study with the DWP, in which I investigated barriers to and opportunities for public sector workers to implement algorithmic bias mitigation techniques. In this paper, I focused on the following

research questions: RQ3a: What aspects of DWP organisational culture might influence the adoption of mitigation approaches? and RQ3b: And, what does this mean for what might work to mitigate algorithmic bias in practice? The data for this paper were collected through conducting a series of seven educational workshops on algorithmic bias mitigation, and seven follow up interviews with practitioners in the UK government department, The Department of Work and Pensions (DWP). The workshops focused on how algorithmic bias might develop, and explored bias mitigation tools such as algorithmic impact assessments and value sensitive design. After these workshops participants were invited to take part in a follow up interview, to allow them to reflect on the content of the workshops and its relevance to their working practices.

This third paper identified three key findings, presented as two challenges and one opportunity. The first challenge is that it is difficult for civil service practitioners to align technologies to the social justice values which underline socio-technical bias mitigation approaches when servicing a large diverse public. Participants explored this issue by talking about the rights and perceived expectations of taxpayers, and how taxpayers often had diverse and conflicting views. Furthermore, civil service practitioners' room for action is limited by the political structures they work within, and government policy approaches may sometimes be in opposition to social justice values. The second challenge is that practitioners perceived a lack of clarity within organisational guidance and anti-discrimination legislation about when it was fair to treat different groups differently, which can lead to additional uncertainty over how conflicting needs within the population should be addressed. The opportunity identified in this paper is that participants perceived diversity in the workforce as important to algorithmic bias mitigation efforts. However, due to influences at the organisational level, some participants were uncertain as to how effective this might be.

In the following section, I turn to the general conclusions which can be drawn from these research papers.

7.2. Overarching contributions

In addition to the findings suggested by each of my three empirical papers, some general conclusions have been drawn from analysing these papers together. To consider the overarching contributions of the thesis, I return to its overarching aim: to investigate how the DWP might mitigate the risks of algorithmic bias within their data science team. This aim focuses on the practical implementation of algorithmic bias mitigation methods, and specifically, on the DWP context. Therefore, in analysing my three research papers, it was important to consider both the practical and contextual aspects of algorithmic bias mitigation.

Three overarching contributions and one methodological contribution emerged from this analytical process. The first overarching contribution*, Time management and project deliverables* focuses on how the type of fast paced working practices found in the development of algorithmic technologies is not conducive to the type of slower paced thinking needed to consider the issue of algorithmic bias. The second, *The significance of the civil service context*, focuses on the challenges presented by the civil service context. The third, *The missing public* addresses how while practitioners' perceptions of the public were

ever present during the interviews, the voice of the public themselves was missing. In addition to these overarching contributions, I make a methodological contribution in the form of a reflection on the use of workshops as a data collection method. I expand on these below.

Time management and project deliverables

In the findings of papers one and two, participants described fast paced agile working practices. I suggest that these papers show that these types of working practices are not conducive to the slower type of thinking required to consider the issue of algorithmic bias. In the first paper, DWP data science practitioners worked in a team which took on work from other DWP teams. These other teams would be the project's clients. Participants described how the turnaround time on projects was rapid and relied on agile-styled project management working practices. Additionally, participants stressed the importance of practitioners understanding their remit on the project to further expedite efficient working practices. In the second paper, a participant on the Aurora AI ethics committee described how he felt the ethics committee's concerns regarding algorithmic bias had been sidelined due to delivery pressures on the Aurora AI project. Furthermore, one of the project's senior advisors described how he expected the Aurora AI ethics committee to be in alignment with the aims of the project and not impede the project's development. The working practices and organisational culture described in these papers suggest typical project management styles found in technological development may not be conducive to the type of thinking required to consider wider issues regarding injustice.

When studying the ways in which AI practitioners attribute ethical responsibility, Orr and Davies (2019) found that due to limited time, data scientists were unable to use what they perceived as the most thorough quantitative methods when developing an algorithmic technology. Additionally, they found that the client-customer relationship they had in the context of their work limited the approaches available to them, which echoes the findings from my first paper. Furthermore, the findings presented in my research papers advance this observation, finding not only that organisational culture limits the time available to use more thorough quantitative methods, but also that the fast-paced culture and agile project management style encouraged data scientists to prioritise delivering on the project's technical outcomes over considerations of algorithmic bias mitigation. This system of prioritisation and its accompanying time pressure did not appear to provide practitioners with the time for reflective contemplation regarding the technologies' potential biases from a socio-technical standpoint. In other words, it not only limits practitioners' methodological choice regarding how to develop algorithmic technologies, but also limits thinking which is not geared towards delivering the product on time.

Whilst it could be argued that allowing practitioners more time to develop algorithmic technologies, and de-prioritising the completion of a finished product, will not address issues of algorithmic bias by itself, I suggest this would provide a starting point for practitioners to discuss issues of algorithmic bias using a socio-technical lens. As discussed in Chapters 2 and 4, much of the critique surrounding technical de-biasing methods is rooted in the 'framing

trap' (Selbst *et al.*, 2018). To move beyond a data and algorithm-based framing of algorithmic bias mitigation, practitioners will need to have the time to consider approaches and methods which are outside of their typical approach. This suggests that organisational working practices which emphasise the need to deliver unbiased algorithmic technologies, alongside sufficient time to consider algorithmic bias from a socio-technical perspective, needs to be embedded within the digital development cycle.

Agile project management working practices which aim to incorporate and prioritise consideration of the potential harms of algorithmic technologies have been put forward. Moss and Metcalf's (2020) research on Silicon Valley ethics workers reported that some companies have considered how agile working methods might be adjusted to create more ethical working practices. One of these adjustments includes incorporating a component called 'consequence scanning' into the process. However, while this provides a mechanism for reflecting on the potential harms caused by a project during the development process, it does not address the issue of time pressures. As found within my research, practitioners often have mechanisms aimed at incorporating ethical insight (ethics applications, ethics committees, etc.), however, the effectiveness of these methods is hampered by the pressure to quickly deliver tangible results.

The significance of the civil service context

In the findings presented in papers one and three, participants described the challenges present in mitigating algorithmic bias within a civil service context. I suggest that these papers show that challenges are presented by the civil sector context regarding algorithmic bias mitigation. Specifically, these challenges are caused by the influence of political processes within this context. In the first of my research papers, DWP participants described how pre-existing organisational knowledge production limited what approaches might be considered to mitigate algorithmic bias. In particular, participants described how the team was expected to incorporate the department's previous research findings regarding the success of conditionality into the Digital Plus Trial project. In my third research paper, DWP participants described the difficulty for practitioners to align technologies to social justice values when servicing a large heterogenous public, due to the political structures within which DWP is embedded. Furthermore, these structures restricted how practitioners might engage with social justice concepts, due to the political influence of the government and political lobbying groups.

In paper one, my findings showed participants needed to ensure that welfare conditionality persisted throughout the Digital Plus Trial. Rodger (2012) argues that welfare systems which enforce conditionality allow the state to criminalise benefit claimants, by attempting to compel claimants into behaving in a specific way to qualify for assistance which gives them access to necessary material goods, such as food and housing (Rodger, 2012). Furthermore, he argues that conditionality is ideologically aligned with systems that support discourses framing those in receipt of benefits as being either the 'deserving' or 'undeserving' poor. As Eubanks highlights in *Automating Inequality*, it is these discourses that serve to individualise

the struggles of those in poverty and minimise their framing as a societal issue (Eubanks, 2018). Thus, the political environment in which the civil service develops algorithmic technologies, and the power relations which exist therein, is of particular interest in assessing the potential harms caused by these technologies.

The findings in paper three are particularly relevant to the challenges in addressing algorithmic bias in the public sector and the civil service. In Chapter 3 it was argued that data scientists must adopt social justice-based approaches in order to address algorithmic bias (Green, 2021). However, practitioners may perceive themselves as having a duty to respect the values of all citizens, not only those whose values align with those of social justice. Research from the *Living With Data* project investigating public perception regarding the use of data sharing in three public sector contexts, stresses the need to recognise there is not a single 'public', but diverse publics (Ditchfield *et al.*, 2022). This research demonstrated that diverse publics perceived data uses differently – they had diverse concerns, expectations, and envisioned different possibilities and risks regarding the use of their data.

These findings suggest that the civil service context has particular challenges in regard to algorithmic bias mitigation. Firstly, algorithmic technologies deployed in this context are strongly influenced by political processes, and build on policy decisions already put in place by government officials. Second, due to data practitioners' roles as civil servants in this context, these practitioners may be required to balance the views of diverse publics in a way private organisations do not. Since these findings result from the embedding of civil service institutions in the wider political landscape, they are likely generalisable to other UK civil service contexts.

The missing public

In the findings presented in all three papers, participants described their difficulties in attempting to include the voice of the public in algorithmic bias mitigation processes. In paper one, participants described how they had engaged the public regarding the DTF and Digital Plus Trial. However, this engagement was primarily to elicit claimants' perceptions of the trial, rather than participants' feelings regarding concerns about the possibility of bias arising from their data use in the trial. Furthermore, this paper presented the finding that although participants on the DTF and Digital Plus Trial felt a strong sense of responsibility to the general public, they were directly accountable to the government, and not the public. In paper two, participants described how they had attempted to use co-design principles in the development of the Aurora AI project. However, during these attempts, only a very narrow range of citizens were engaged with about their thoughts and feelings on the project. These citizens were, for the most part, from more privileged backgrounds, and it is unlikely they would be able to comment on the experiences of people from other marginalised backgrounds. Furthermore, the manner in which these data were collected more strongly resembled user feedback. In paper three, participants explained how it was difficult to assess public perception, and when faced with diverse views, it was difficult to know whose voice should be prioritised.

For socio-technical algorithmic bias mitigation methods such as AIAs to be effective, it is important they include the voice of the public to assist in co-constructing the potential harms and impacts posed by these technologies (Metcalf *et al.*, 2021). Without this input, it is difficult to know whether the potential harms documented in these assessments reflect the publics' concerns regarding these technologies. The Ada Lovelace Institute's report from their work piloting an AIA for an NHS context (2022), stresses the need for the public to participate in these assessments. The report also highlights how participatory mechanisms for involving the public will likely be different depending on the context within which the algorithmic technology will be deployed. These findings suggest the need for further research in incorporating the missing public into the discussion of algorithmic bias mitigation. However, exploring the effectiveness of different participatory mechanisms for this purpose is beyond the scope of my research.

Reflections workshops for data collection

In my third paper, I used educational workshops as a data collection method. This served two purposes beyond data collection. First, it allowed me to introduce participants to the socio-technical algorithmic bias mitigation methods which emerged from the findings of the second paper in an environment characteristic of DWP working practices. Workshops also allowed me to bring together participants from different backgrounds, based on the finding from my second research paper that practitioners needed to create a shared understanding of algorithmic bias. Secondly, using workshops as a research method gave me the opportunity to strengthen the partner relationship between myself and the DWP by providing value in return for their support early on, rather than waiting for the project's research outputs. I expand on these points below.

To discuss the benefits of using this method, it is important to return to the overarching aim of my thesis; to investigate how the DWP might mitigate the risks of algorithmic bias within their data science team. The aim of my thesis was not only to address research questions of academic interest, but also to provide insight into algorithmic bias mitigation for use in a public sector organisation. As discussed in Chapter 3, Blaikie (2005) describes this type of research as "applied research" – research which typically focuses on change, evaluation, or assessing social impacts. The type of applied research most relevant to my third research paper was 'change'. Blaikie (2005) describes research focusing on change as "[intervening] in a social situation by manipulating some aspects of it, or to assist the participants in doing so, preferably on the basis of established understanding or explanation."

For my thesis to provide a 'change' outcome, it was necessary for my third paper to include an intervention based on the findings of paper two. This intervention was split into two parts. First, it introduced participants to the socio-technical algorithmic bias mitigation methods which emerged from the findings of the second paper. Second, as the findings of the second research paper suggested that mitigating algorithmic bias is challenged by practitioners lacking a shared sense of understanding of algorithmic bias, the workshops aimed to bring

together practitioners from different backgrounds, departments, and roles, to discuss algorithmic bias together and gain that sense of shared understanding.

As discussed in Chapter 3, workshops, like focus groups, include interaction between participants, and allow opinions to be revealed which might not otherwise have surfaced in a traditional one to one interview (Morgan, 1998). This echoes my experience using workshops as a research method. During my follow up interviews, participants mentioned one of the things they had found insightful during the workshops was speaking to colleagues in other segments of the organisation. They stated themselves these participants were often working on things they that had not been aware of, and they found this aspect of the workshops particularly valuable. Reflecting on my experience running these workshops, they provided participants with a space to discuss these issues, and to form relationships throughout the organisation. Whilst analysing the workshop transcripts from the breakout rooms, it was interesting to see different participants take on different roles in the group. Some of the more 'techy' participants would take on the role explaining how algorithms worked, and answer the less 'techy' participants questions about algorithms. Other participants would share their experiences of working on diversity initiatives and how they felt this related to algorithmic bias. During brainstorming activities, some participants would take on the role of the 'ideas person', while others would be the 'details person'. Throughout the workshops and the workshop activities, it was interesting to see how people came together to focus on these tasks. This gave insight not only into organisational practices at the DWP, but also how practitioners from different backgrounds might think about these issues and communicate about them with others.

Regarding the second purpose, the workshops allowed me to provide my partner organization value with respect to their continued support during my PhD research. As these workshops were designed to be educational, they provided the DWP with value from the research partnership prior to the finished research outputs, and strengthen the relationship between myself and DWP contacts. Mason and Siddique (2023), reflecting on the challenges they have experienced during academic-community research collaborations, argue that social science research is often extractive and rarely provides value for the partnership organisation. Additionally, they proposed that academics often do not provide their partner organisation with value or insights in a timely manner. Failure to provide value in these partnerships may put stress on the academic-organisation relationship. This can be a prominent issue if researchers, in the course of their research, make frequent requests on organisation time and resources. In using workshops, I combined both my own and my DWP contacts' goals regarding the research. I was able to investigate DWP practitioners' perceptions of algorithmic bias mitigation methods, and they received workplace training on issues of algorithmic bias.

By combining these aims, DWP practitioners were able to see value in the research I was doing within the organisation. This was demonstrated by a participant approaching me after they had attended some of the original six workshops, to ask me about conducting another workshop in a different segment of the organisation, which suggests the DWP found these workshops valuable in their approach to algorithmic bias mitigation. Following my PhD

research, I am in discussion with the DWP about conducting further research and knowledge exchange events based on the workshop content I used in my third paper. This may also suggest that by providing value to organisations throughout the course of research, following through with one's promises, and building a steady relationship built on trust, participants may be able to assist researchers in accessing other areas of the organisation in question.

However, in hindsight, the workshop format also presented its own challenges. I designed the first six workshops as a series covering different topics related to algorithmic bias mitigation, as I wanted to thoroughly cover each topic. This plan was not well suited to the DWP organisational context - the original series of six workshops proved too long for participants to be able to continually attend alongside their other work commitments. Although the first workshop of the original six was attended by eighteen participants, this reduced to two in the final two workshops. As the workshops were designed to lead on from one another, this presented particular barriers during the research process. Participants who joined the series halfway through were given the slides from previous lectures, and links to the video recording of the session on the DWP's system. In retrospect, it would have been beneficial to reduce the number of workshops to ensure they fitted easily into my participants' schedules. My experience echoes Mason and Siddique's (2023) comments regarding how time intensive it can be for organisations to work with academics and suggesting the need for researchers to adjust their approach to stay respectful of the organisations who are collaborating with them.

7.3. Recommendations for practice

From my papers, I identify three recommendations for practice:

(1) Practitioners from different disciplines and roles need to create a shared understanding of algorithmic bias, and
(2) UK civil servants seeking to further social justice aims must scope out potential routes that are possible within the constraints of the civil service, and
(3) Practitioners should seek to adopt socio-technical algorithmic bias mitigation methods, such as AIAs, VSD, and critical thinking about data and their wider environment.

I expand on these below:

The first of my recommendations is that effort is required to create a shared understanding of algorithmic bias across stakeholders with often disparate skillsets, expertise, knowledge-bases, values, and beliefs. Across the three chapters, participants operationalised different definitions of bias and discrimination, with more technical participants focusing on accessibility or legal frameworks. In the second of my research papers, these differences in understanding, combined with systemic issues such as funding and organisational structures, meant the Aurora AI ethics committee felt ethical concerns were sidelined. Thus, attempts to mitigate bias can be hampered by a lack of a shared understanding of core concepts and of the potential impacts of these technologies. This may be achieved by ensuring projects include a range of interdisciplinary perspectives, and those working on a project are

encouraged to discuss any differences which may exist between what can seem like a shared vocabulary.

The second of my recommendations is that civil servants wishing to further social justice aims in mitigating the risks of algorithmic bias may need to scope out the possibilities for potential action within their own organisational contexts. This results from two unique challenges presented by the UK civil service sector. One, the political structures within which civil service departments are embedded restrict how practitioners might engage with social justice concepts, due to the political influence of the government and lobbying groups. Two, the stipulation that civil servants are publicly apolitical may limit the routes available for practitioners to protest against unjust technological outcomes in this context. Thus, when attempting to further social justice aims, civil servants will need to identify what routes are available to them to effectively protest against political influences which may lead to algorithmic bias. Specifically, this could mean addressing the following questions; what does centring marginalised groups look like in a civil service context? How might this be at odds with the structures it is embedded within and relied upon? What other limitations might practitioners face in this space?

The third of my recommendations is the use of socio-technical algorithmic bias mitigation methods such as VSD, AIAs, and critical thinking around datasets. During my first and third research papers, participants found established processes for ethical decision making crucial in the context of their work. Additionally, during the third of my research papers, participants found the AIA workshop to be the most useful and relevant to their everyday working practices. Whilst participants did not find the VSD and critical thinking exercises as useful from a working practices perspective, they still found these sessions interesting. I recommend public sector organisations intending to mitigate the impact of algorithmic bias should seek to incorporate socio-technical algorithmic bias methods in a way which matches the working context. Furthermore, methods which resemble current working practices may be more easily adopted.

7.4. Theoretical contribution: A meso-level intervention

With the overarching contributions of my thesis in mind, I now turn to how these, and the empirical contributions of each paper, form the basis of the theoretical contribution of my thesis – furthering the study of algorithmic bias from a meso-level of analysis in the UK civil service sector. Additionally, I suggest the importance of furthering the study of this level of analysis in other sectors, due to the likelihood of meso-level social dynamics which are context specific.

To date, much of the research on algorithmic bias has focused on developing an understanding of the micro-level (actions of individuals) and macro-level (political and societal influences) factors involved in the development of algorithmic bias. Eubanks (2018) investigates numerous algorithmic projects, including algorithms designed to assess the risk of a child being abused, housing support algorithms, and more general welfare provision

algorithms. She describes the connection between the actions of individuals, both those who use these algorithmic technologies and those subject to their decisions, and political ideologies such as neoliberalism. Furthermore, she argues that neoliberal values become embedded in algorithmic technologies due to the decisions made during the development of algorithmic technologies, which are influenced by political agenda setting. However, whilst she describes many factors of this phenomenon at both a micro and macro-level, less attention is paid to the organisational and project level dynamics which mediate the influence of macro-level factors. Or, to put it another way, less attention is paid to how organisations may interpret and implement these policies, and what other factors organisation's may be influenced by during this process.

Similarly, O'Neil (2017) investigates several algorithmic projects from the perspective of a data scientist. From this perspective, she provides a theory if how the individual worldviews of a data scientist can become embedded within algorithmic technologies. Furthermore, she describes how these micro-level influences are tied to wider injustices and societal prejudices, such as how people are oppressed on the basis of race, class, gender, etc. In a similar vein, Orr and Davies (2020), Veale et al. (2018), and Holstein (2017) have interviewed practitioners to understand how they are situated within the development of algorithmic bias, and their responsibilities and engagements within their working context. However, their analysis is primarily focused on how individual actors are constrained within a collective system (Holstein, McLaren and Aleven, 2017; Veale, Van Kleek and Binns, 2018; Orr and Davis, 2020).

My research attempts to extend the meso-level of study of algorithmic bias by suggesting the importance of the organisational context, in addition to suggesting factors and dynamics which might be saliant to the UK civil sector. Furthermore, I identify organisational working practices as important factors for consideration, such as project management styles, organisational guidance, and organisational perceptions of responsibility and accountability within a complex environment. I expand on this below.

My research highlights the importance of the organisational context when analysing algorithmic bias development and mitigation efforts. As discussed in my overarching contributions, the findings in my research papers suggest that the UK civil service context has unique challenges in regard to algorithmic bias mitigation. Firstly, algorithmic technologies deployed in this context are strongly influenced by political processes and build on policy decisions already put in place by government officials. Second, due to data practitioners' roles as civil servants, these practitioners may be required to balance the views of diverse publics in a way private organisations do not. This is due to the requirement that civil servants must demonstrate political impartiality and must not allow their political position to determine any advice they may give or their actions (Civil Servants, Ministers and Parliament, n.d.). This suggests the importance of considering organisation and sector specific guidance as valuable factors of study in the development of algorithmic bias mitigation theory. Extrapolating from this, one may assume, each sector will have its own challenges. I therefore argue that it is important to develop a strong meso-level understanding of different sectors approaches, challenges, and working practices in relation

to algorithmic bias. This is important in order to uncover sector specific factors which may help or hinder algorithmic bias mitigation efforts.

Furthermore, as suggested by the findings in my third research paper, it is important to recognise that organisational culture may influence practitioners' perception and interpretation of laws related to algorithmic bias mitigation, and this may shape practitioner's working practices. Within the context of my third research paper, practitioners perceived that regulatory frameworks require treatment to be equal, and so their work requires, at least in part, to pay careful attention to statistical and procedural conceptualisations of fairness. This builds on previous theory regarding algorithmic bias mitigation, where critiques of statistical and procedural fairness have rightfully come to the fore (Green and Hu, 2018; Hoffmann, 2019), and suggests organisational factors which influence practitioners' perceptions of regulatory frameworks may provide obstacles to moving beyond these types of approach. Furthermore, while organisations are required to uphold the law, there is little research in the field of critical algorithm studies into whether the interpretation and implementation of the law differs between organisations and sector specific contexts. This highlights the importance of meso-level analysis, as differences in organisation or sector interpretation of the law or other types of algorithmic bias mitigation guidance may help or hinder algorithmic bias mitigation processes.

Alongside these factors, I also argue that organisational working practices such as project management styles, organisational guidance, and organisational perceptions of responsibility and accountability within a complex environment are important. As discussed by Moss and Metcalf (2020), Silicon Valley 'ethics workers' are often constrained by the organisational values, organisational structure, and working practices they are embedded within. This is supported by Orr and Davies (2019), who found that due to limited time, data scientists were unable to use what they perceived as the most thorough quantitative methods when developing an algorithmic technology. The findings presented in my research papers advance this observation, finding not only that organisational culture limits the time available to use more thorough quantitative methods, but also that the fast-paced culture and agile project management style encouraged data scientists to prioritise delivering on the project's technical outcomes over considerations of algorithmic bias mitigation. This system of prioritisation and its accompanying time pressure did not appear to provide practitioners with the time for reflective contemplation regarding the technologies' potential biases from a socio-technical standpoint. I suggest these findings support my argument for the importance of meso-level analysis in the development of algorithmic bias and algorithmic bias mitigation theory.

7.5. Limitations

Each of the chapters forming this thesis has its own limitations. In Chapter 4, the fieldwork was greatly impacted by the outbreak of Covid-19, and therefore its methodology had to be

altered from its original plan. The original methodology planned for the fieldwork to occur in two stages. In the first stage, I would conduct scoping interviews with key members of the DTF project, and in the second stage, I would conduct code-based elicitation interviews with data scientists. These interviews would have involved data scientist participants showing me code they had developed on the DTF and Digital Plus Trial as an elicitation device to understand their working practices. To compensate for the interruption of Covid-19, changes were made to the methodology to expand the number of first stage interview participants, to interview some participants twice. However, the small team size on the Digital Trialling Framework project, this limited the available sample. Because of the small sample size in this case study, its findings may lack transferability to other contexts. In addition, as some of the interviews were conducted during the pandemic – and after the DTF project had been cancelled – participants may not have remembered as much as they would have without the interruption.

The Digital Plus Trial was being designed to investigate whether work coaches could save time by engaging with the more 'digitally competent' claimants with 'less complex needs' online, instead of seeing them in a face-to-face setting. Prior to the outbreak of Covid-19, the DWP primarily engaged with claimants face-to-face, during which time the proposed Digital Plus Trial was considered controversial in the department. This was because The Treasury was concerned that by moving the service online, claimants would not be subject to the same level of conditionality as in a face-to-face setting. However, due to the outbreak of Covid-19, DWP were required to move all claimants' work coach meetings online. Consequently, participants interviewed after the outbreak of Covid-19 may have been less critical of the Digital Plus Trial than they were previously, following the department's recent shift towards digital engagement.

In Chapter 5, six of the participants had worked on the Aurora AI project and seven were AI Ethics experts. In total, 13 participants were interviewed during this study, which can be considered a small sample size. Furthermore, of the Aurora AI participants, all had different roles on the project. While a qualitative study does not aim to be generalisable, it would have been beneficial to have interviewed more people on the Aurora AI project – particularly, with the sample having multiple people with similar roles on the project, to improve the robustness of the findings. Another important limitation on this project resulted from the fact that I only speak in English, and therefore interviews had to be conducted in English. On the Aurora AI side of my sample, some participants were less comfortable than others speaking in English, which will likely have led me to missing some of the linguistic and cultural nuance in these exchanges.

Finally, as applied research using qualitative methods, it is important to note that the findings of my papers may not be generalisable to other contexts. Whilst there will likely be some generalisability to other UK civil service contexts, due to the similarity of the organisations and the way they are embedded within UK political structures, further research is required to better understand the differences which might exist between civil service organisations in the UK. In addition, comparative research with other countries' civil service organisations would be beneficial.

7.6. Further research

My findings suggest an avenue for further research into the relationship between human and algorithmic decision making.

Direct comparisons between human and algorithmic decision-making processes have been scarce. In paper three, participants described their uncertainty over whether humans defaulted to algorithmically generated decisions in their decision-making processes, as suggested by previous research (e.g., Eubanks). Instead, participants suggested decision makers may default to recommendations provided to them by either algorithmic technologies or human decision makers. Participants made this observation using their experiences with human recommendations provided to the department from a third sector organisation when conducting health assessments. These recommendations went on to inform whether a claimant would receive health related benefits. Participants noted that in these cases, they would often defer to the third sector organisation's recommendation. Participants also described how they were aware of cases in the department in which humans did not follow machine generated recommendations. This suggests a need to further research the complex interplay of human and algorithmically derived decisions and their consequences, in specific contexts.

Additionally, my findings suggest the need for further research regarding the implementation of socio-technical methods such as VSD and AIAs. While practitioner perceptions in my research suggest these may be promising for the civil service context, there is still little research which has implemented these mechanisms in practice. Furthermore, my thesis suggests the need for research which focuses on how practitioners might incorporate the public's voice into the process of using either VSD or AIA methods. The Ada Lovelace Institute has started this work with the NHS Imagining library (Ada Lovelace Institute, 2022), and the participatory design process they have developed as part of it. However, more research is needed to understand how these mechanisms differ in different organisational contexts and fit into established working practices.

7.7. Concluding remarks

Mitigating the effects of algorithmic bias completely will be very difficult to achieve. Although my research investigated how the DWP might seek to mitigate algorithmic bias, I am aware that my findings appear to have unearthed more challenges than solutions to this problem. This is partly due to my own orientation to the problem; I am researching this because I care about social justice and discrimination, and I am critical of technology uses which result in discrimination. However, it is important to note the positives which came out of this research process, in particular the involvement and engagement of the DWP in supporting this PhD.

As mentioned above, during the research for paper three, I was contacted by a data scientist about giving an extra workshop because of the interest in this topic in the cyber security team. After sending out invitations, word spread to other segments of the organisation, and I had

more emails about whether people could drop in. My participants were incredibly engaged during the event, and during the follow up interviews many expressed a wish to know more about algorithmic bias and algorithmic bias mitigation methods.

Additionally, the research partnership upon which this PhD has been built has been extremely positive. Since finishing my research, my contacts at the DWP have expressed interest in the content of my workshops being suitably packaged for the DWP intranet, so others in the organisation can access the same resources. It has also been suggested that I return to provide workshops for different teams. Furthermore, my research on algorithmic bias mitigation and AIAs was included in a recent DWP submission to the Cabinet Office on how the department is ensuring that it uses AI responsibly. While mitigating algorithmic bias is incredibly complex, challenging, and requires co-ordination between many different actors, my experience during my PhD suggests there is substantial interest in meeting these challenges.

Bibliography

Acas (2023) *Using protected characteristics to make decisions*, *Acas*. Available at: https://www.acas.org.uk/employer-decision-protected-characteristic (Accessed: 5 October 2023).

ACM (no date) *The Code affirms an obligation of computing professionals to use their skills for the benefit of society.* Available at: https://www.acm.org/code-of-ethics (Accessed: 11 August 2023).

Ada Lovelace Institute (2022) *Algorithmic impact assessment: a case study in healthcare*. Ada Lovelace Institute, pp. 1–119. Available at: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/.

Ahmed, S. (2012) *On being included + Racism and Diversity in Institutional Life*. Available at: https://books.google.co.uk/books?id=7HKk_DLzRDAC&printsec=frontcover&dq=ahmed+on+being+included&hl=en&sa=X&redir_esc=y#v=onepage&q=ahmed on being included&f=false.

Alibhai, Z. (2019) *DWP reportedly bans Jobcentre referring benefit claimants to foodbanks*, *inews.co.uk*. Available at: https://inews.co.uk/news/uk/universal-credit-dwp-jobcentre-food-banks-ban-claims-256581 (Accessed: 23 August 2023).

Ananny, M. and Crawford, K. (2018) 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability', *New Media & Society*, 20(3), pp. 973–989. Available at: https://doi.org/10.1177/1461444816676645.

Angwin, J. *et al.* (2016) 'Machine Bias — ProPublica', *ProPublica*. Available at: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (Accessed: 31 October 2019).

Ashley, L. (2010) 'Making a difference? The use (and abuse) of diversity management at the UK's elite law firms', *Work, Employment and Society*, 24(4), pp. 711–727. Available at: https://doi.org/10.1177/0950017010380639.

Baker, M., French, E. and Ali, M. (2021) *Insights into Ineffectiveness of Gender Equality and Diversity Initiatives in Project-Based Organizations*.

Balayn, A. and Gürses, S. (2021) *Beyond de-biasing: Regulating AI and its inequalities*. Belgium: European Digital Rights (ESRi). Available at: https://edri.org/our-work/if-ai-is-the-problem-is-debiasing-the-solution/.

Bates, J. *et al.* (2023) 'Socially meaningful transparency in data-based systems: reflections and proposals from practice', *Journal of Documentation*, ahead-of-print(ahead-of-print). Available at: https://doi.org/10.1108/JD-01-2023-0006.

BBC News (2021) 'Universal credit fraud and error at new high', 13 May. Available at: https://www.bbc.com/news/uk-57091551 (Accessed: 23 August 2023).

Bellamy, R.K.E. *et al.* (2019) 'AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias', *IBM Journal of Research and Development*, 63(4/5), p. 4:1-4:15. Available at: https://doi.org/10.1147/JRD.2019.2942287.

Bender, E.M. *et al.* (2021) 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery (FAccT '21), pp. 610–623. Available at: https://doi.org/10.1145/3442188.3445922.

Big Brother Watch (2021) *Poverty Panopticon: the hidden algorithms shaping Britain's welfare state*. Big Brother Watch.

Big Brother Watch Team (2018) *A Closer Look at Experian Big Data and Artificial Intelligence in Durham Police — Big Brother Watch*, *A Closer Look at Experian Big Data and Artificial Intelligence in Durham Police — Big Brother Watch*. Available at: https://bigbrotherwatch.org.uk/2018/04/a-closer-look-at-experian-big-data-and-artificial-intelligence-in-durham-police/ (Accessed: 6 August 2023).

Binns, R. (2018) 'Algorithmic Accountability and Public Reason', *Philosophy and Technology*, 31(4), pp. 543–556. Available at: https://doi.org/10.1007/s13347-017-0263-5.

Bogner, A., Littig, B. and Menz, W. (2018) 'Generating Qualitative Data with Experts and Elites In: The SAGE Handbook of Qualitative Data Collection'. Available at: https://doi.org/10.4135/9781526416070.

Bostrom, N. (2014) 'Superintelligence: Paths, dangers, strategies.', *Superintelligence: Paths, dangers, strategies.* [Preprint]. Available at: https://doi.org/10.1017/S0031819115000340.

Braun, Clarke, and Hayfield (no date) *Qualitative Psychology: A Practical Guide to Research Methods*. Available at: https://books.google.co.uk/books?hl=en&lr=&id=lv0aCAAAQBAJ&oi=fnd&pg=PA222&dq=braun+and+clarke+thematic+analysis&ots=eOLGhtdlLB&sig=omDxEvzuvz2dRujqrvLiEOjJjvo&redir_esc=y#v=onepage&q=braun and clarke thematic analysis&f=false (Accessed: 25 August 2022).

Braun, V. and Clarke, V. (2006) 'Using thematic analysis in psychology', *Qualitative Research in Psychology*, 3(2), pp. 77–101. Available at: https://doi.org/10.1191/1478088706qp063oa.

Braun, V. and Clarke, V. (2021) 'One size fits all? What counts as quality practice in (reflexive) thematic analysis?', *Qualitative Research in Psychology*, 18(3), pp. 328–352. Available at: https://doi.org/10.1080/14780887.2020.1769238.

Breiman, L. (2001) *Statistical Modeling: The Two Cultures*, *Statistical Science*, pp. 199–231.

Bryman, A. (2012) *Social research methods*. Oxford: Oxford University Press.

Buchanan, B.G. *et al.* (1976) 'Applications of artificial intelligence for chemical inference. 22. Automatic rule formation in mass spectrometry by means of the meta-DENDRAL program',

Journal of the American Chemical Society, 98(20), pp. 6168–6178. Available at: https://doi.org/10.1021/ja00436a017.

Buchter, L. (2021) 'Escaping the Ellipsis of Diversity: Insider Activists' Use of Implementation Resources to Influence Organization Policy', *Administrative Science Quarterly*, 66(2), pp. 521–565. Available at: https://doi.org/10.1177/0001839220963633.

Butler, P. and editor, P.B.S. policy (2022) 'DWP blocks data for study of whether benefit sanctions linked to suicide', *The Guardian*, 2 March. Available at: https://www.theguardian.com/society/2022/mar/02/dwp-blocks-data-for-study-of-whether-benefit-sanctions-linked-to-suicide (Accessed: 23 August 2023).

Campolo, A. *et al.* (2017) *AI Now 2017 Report*. AI NOW.

Carbado, D.W. and Gulati, M. (2000) 'Working Identity', *CORNELL LAW REVIEW*, 85.

Carlile, P.R. (2002) 'A Pragmatic View of Knowledge and Boundaries: Boundary Objects in New Product Development', *Organization Science*, 13(4), pp. 442–455. Available at: https://doi.org/10.1287/orsc.13.4.442.2953.

Clement-Jones, T. (2021) 'Tackling the algorithm in the public sector', *The Constitution Society*, 19 March. Available at: https://consoc.org.uk/tackling-the-algorithm-in-the-public-sector/ (Accessed: 15 January 2023).

Costanza-Chock, S. (2018) 'Design Justice, A.I., and Escape from the Matrix of Domination', *Journal of Design and Science* [Preprint]. Available at: https://doi.org/10.21428/96c8d426.

Costanza-Chock, S. (2020) *Design Justice*.

Cramer, H. *et al.* (2018) 'Assessing and addressing algorithmic bias in practice', *Interactions*, 25(6), pp. 58–63. Available at: https://doi.org/10.1145/3278156.

Cristi, C. (2019) *L.A.'s homeless: Aerial tour of Skid Row, epicenter of crisis*, *ABC7 Los Angeles*. Available at: https://abc7.com/homeless-homelessness-los-angeles-la/5344680/ (Accessed: 22 September 2023).

Danks, D. and London, A.J. (2017) 'Algorithmic Bias in Autonomous Systems A Taxonomy of Algorithmic Bias', *26th International Joint Conference on Artificial Intelligence (IJCAI-17)*, (Ijcai), pp. 4691–4697.

Data Ethics Framework (no date) *GOV.UK*. Available at: https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-2020 (Accessed: 30 April 2023).

Daws, R. (2020) *Experts believe Industry 4.0 will drive a century of advancements over five years*, *Internet of Things News*. Available at: https://www.iottechnews.com/news/2020/sep/25/experts-industry-4-drive-century-advancements-five-years/ (Accessed: 18 July 2023).

Delmont, S. and Mason, J. (1997) 'Qualitative Researching', *The British Journal of Sociology* [Preprint]. Available at: https://doi.org/10.2307/591613.

Dencik, L. *et al.* (2018) *Investigating uses of citizen scoring in public services*. Data Justice Lab.

Dencik, L. *et al.* (2022) *Data Justice*. London, UNITED KINGDOM: SAGE Publications, Limited. Available at: http://ebookcentral.proquest.com/lib/sheffield/detail.action?docID=7121001 (Accessed: 23 August 2023).

Dencik, L., Hintz, A. and Cable, J. (2016) 'Towards data justice? The ambiguity of anti-surveillance resistance in political activism', *Big Data & Society*, 3(2), p. 2053951716679678. Available at: https://doi.org/10.1177/2053951716679678.

Dictionary.com (no date) *Definition of fair | Dictionary.com*, *www.dictionary.com*. Available at: https://www.dictionary.com/browse/fair (Accessed: 8 May 2023).

D'Ignazio, C. and Klein, L.F. (2020) *Data Feminism*, *Data Feminism*. Available at: https://doi.org/10.7551/mitpress/11805.001.0001.

Ditchfield, H. *et al.* (2022) 'Report on Living With Data Interviews & Focus Groups'.

Dobbin, F. and Kalev, A. (2016) 'Why Diversity Programs Fail', p. 10.

Dressel, J. and Farid, H. (2021) 'The Dangers of Risk Prediction in the Criminal Justice System', *MIT Case Studies in Social and Ethical Responsibilities of Computing* [Preprint], (Winter 2021). Available at: https://doi.org/10.21428/2c646de5.f5896f9f.

DWP (2023) *Annual Report and Accounts 2022-23 for the year ended 31 March 2023*. United Kingdom: The Department of Work and Pensions.

Dwyer, P. and Wright, S. (2014) 'Universal credit, ubiquitous conditionality and its implications for social citizenship', *Journal of Poverty and Social Justice*, 22(1), pp. 27–35. Available at: https://doi.org/10.1332/175982714X13875305151043.

Edwards, L. and Veale, M. (2017) 'Slave to the Algorithm? Why a Right to Explanationn is Probably Not the Remedy You are Looking for', *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.2972855.

Eubanks, V. (2018) *Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, *St. Martin's Press*. New York, NY, USA: St. Martin's Press. Available at: https://doi.org/10.1038/scientificamerican1118-68.

Faghmous, J.H. and Kumar, V. (2014) 'A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science', *Big Data*, 2(3), pp. 155–163. Available at: https://doi.org/10.1089/big.2014.0026.

Ferryman, K. and Pitcan, M. (2018) 'Fairness in Precision Medicine', *Data & Society*, February(February), p. 58.

Freytag, P. and Young, L. (2017) *Collaborative Research Design: Working with Business for Meaningful Findings*. Available at: https://doi.org/10.1007/978-981-10-5008-4_14.

Friedman, B., Hendry, D.G. and Borning, A. (2017) 'A survey of value sensitive design methods', *Foundations and Trends in Human-Computer Interaction*, 11(23), pp. 63–125. Available at: https://doi.org/10.1561/1100000015.

Friedman, B. and Nissenbaum, H. (1996) 'Bias in Computer Systems', *ACM Transactions on Information Systems*, 14(3), pp. 330–347. Available at: https://doi.org/10.1145/230538.230561.

Galhotra, S., Brun, Y. and Meliou, A. (2017) 'Fairness Testing: Testing Software for Discrimination', in *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pp. 498–510. Available at: https://doi.org/10.1145/3106237.3106277.

GDPR (2016) 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)', *https://webarchive.nationalarchives.gov.uk/eu-exit/https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504* [Preprint].

Ghadiri, M., Samadi, S. and Vempala, S. (2021) 'Socially Fair k-Means Clustering', *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 438–448. Available at: https://doi.org/10.1145/3442188.3445906.

Gillespie, T. (2014) '167The Relevance of Algorithms', in *Media Technologies: Essays on Communication, Materiality, and Society*. The MIT Press. Available at: https://doi.org/10.7551/mitpress/9780262525374.003.0009.

GOV.UK (2022) *[Withdrawn] Chapter 6: Working with participants with complex needs and/or additional support requirements*, *GOV.UK*. Available at: https://www.gov.uk/government/publications/work-and-health-programme-provider-guidance/chapter-6-working-with-participants-with-complex-needs-and-or-additional-support-requirements (Accessed: 26 September 2023).

Green, B. (2020) 'Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought'. Available at: https://doi.org/10.1145/3351095.3372840.

Green, B. (2021) 'Data Science as Political Action: Grounding Data Science in a Politics of Justice', *Journal of Social Computing*, pp. 249–265. Available at: https://doi.org/10.23919/jsc.2021.0029.

Green, B. and Hu, L. (2018) *The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning*.

Gros, C. (2020) *LIFE EVENTS IN THE DESIGN OF PUBLIC SERVICES Creating Communities of Service in the Finnish National AuroraAI Program*.

Guinaudeau, B. and Guinaudeau, I. (2022) '(When) do electoral mandates set the agenda? Government capacity and mandate responsiveness in Germany', *European Journal of Political Research*, n/a(n/a). Available at: https://doi.org/10.1111/1475-6765.12557.

Guthman, J. and Brown, S. (2016) 'Whose Life Counts', *Science, Technology, & Human Values*, 41(3), pp. 461–482. Available at: https://doi.org/10.1177/0162243915606804.

Hamraie, A. (2013) 'Designing Collective Access: A Feminist Disability Theory of Universal Design', *Disability Studies Quarterly*, 33(4). Available at: https://doi.org/10.18061/dsq.v33i4.3871.

Hao, K. (2020) 'We read the paper that forced Timnit Gebru out of Google. Here's what it says.', *MIT Technology Review*. Available at: https://www.technologyreview.com/2020/12/04/1013294/google-ai-ethics-research-paper-forced-out-timnit-gebru/ (Accessed: 22 August 2023).

Haraway, D. (1988) 'Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective', *Feminist Studies*, 14(3), p. 575. Available at: https://doi.org/10.2307/3178066.

Hartman, T. *et al.* (2020) 'Public perceptions of good data management: Findings from a UK-based survey', *Big Data & Society*, 7(1), p. 2053951720935616. Available at: https://doi.org/10.1177/2053951720935616.

Hayes-Roth, F., Waterman, D.A. (Donald A. and Lenat, D.B. (1983) *Building expert systems*. Reading, Mass. : Addison-Wesley Pub. Co. Available at: http://archive.org/details/buildingexpertsy00temd (Accessed: 21 July 2023).

Heffer, G. (2022) *Fraud and error in benefits payments hit record £8.6billion last year*, *Mail Online*. Available at: https://www.dailymail.co.uk/news/article-10856965/Fraud-error-benefits-payments-hit-record-8-6billion-year-Covid-surge.html (Accessed: 23 August 2023).

Hoffmann, A.L. (2019) 'Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse', *Information Communication and Society*, 22(7), pp. 900–915. Available at: https://doi.org/10.1080/1369118X.2019.1573912.

Holstein, K., McLaren, B.M. and Aleven, V. (2017) 'Intelligent tutors as teachers' aides', in *Proceedings of the Seventh International Learning Analytics & Knowledge Conference on - LAK '17*. New York, New York, USA: ACM Press, pp. 257–266. Available at: https://doi.org/10.1145/3027385.3027451.

Holter, I.M. and Schwartz-Barcott, D. (1993) 'Action research: what is it? How has it been used and how can it be used in nursing?', *Journal of Advanced Nursing*, 18(2), pp. 298–304. Available at: https://doi.org/10.1046/j.1365-2648.1993.18020298.x.

Hooker, G. and Mentch, L. (2021) 'Bridging Breiman's Brook: From Algorithmic Modeling to Statistical Learning', *Observational Studies*, 7(1), pp. 107–125. Available at: https://doi.org/10.1353/obs.2021.0027.

Hughes, O. (2019) *'More than half' of NHS trusts engaged in AI projects, report suggests*, *Digital Health*. Available at: https://www.digitalhealth.net/2019/12/half-of-nhs-trusts-engaged-in-ai-projects-report-suggests/ (Accessed: 19 December 2019).

ICO (2019) *Data protection impact assessments*. ICO. Available at: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/ (Accessed: 20 June 2019).

Jasanoff, S. and Kim, S.-H. (2009) 'Containing the Atom: Sociotechnical Imaginaries and Nuclear Power in the United States and South Korea', *Minerva*, 47(2), pp. 119–146. Available at: https://doi.org/10.1007/s11024-009-9124-4.

Jaton, F. (2021) 'Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application', *Big Data & Society*, 8(1), p. 205395172110135. Available at: https://doi.org/10.1177/20539517211013569.

Jensen, T., Sandström, J. and Helin, S. (2009a) 'Corporate codes of ethics and the bending of moral space', *Organization*, 16(4), pp. 529–545. Available at: https://doi.org/10.1177/1350508409104507.

Jensen, T., Sandström, J. and Helin, S. (2009b) 'Corporate codes of ethics and the bending of moral space', *Organization*, 16(4), pp. 529–545. Available at: https://doi.org/10.1177/1350508409104507.

Jobin, A., Ienca, M. and Vayena, E. (2019) 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence*, 1(9), pp. 389–399. Available at: https://doi.org/10.1038/s42256-019-0088-2.

Jones, D. *et al.* (2020) 'Characterising the Digital Twin: A systematic literature review', *CIRP Journal of Manufacturing Science and Technology*, 29, pp. 36–52. Available at: https://doi.org/10.1016/j.cirpj.2020.02.002.

Kaun, A., Larsson, A.O. and Masso, A. (2023) 'Automating public administration: citizens' attitudes towards automated decision-making across Estonia, Sweden, and Germany', *Information, Communication & Society*, 0(0), pp. 1–19. Available at: https://doi.org/10.1080/1369118X.2023.2205493.

Keep, M. (2023) *Tax statistics: an overview*. Available at: https://commonslibrary.parliament.uk/research-briefings/cbp-8513/ (Accessed: 23 August 2023).

Krutzinna, J. (2021) 'Simulating (some) individuals in a connected world', *Journal of Medical Ethics*, 47(6), pp. 403–404. Available at: https://doi.org/10.1136/medethics-2021-107447.

Kuhlman, C., Jackson, L. and Chunara, R. (2020) 'No computation without representation: Avoiding data and algorithm biases through diversity'. Available at: http://arxiv.org/abs/2002.11836 (Accessed: 9 August 2022).

Loi, M. and Spielkamp, M. (2021) 'Towards Accountability in the Use of Artificial Intelligence for Public Administrations; Towards Accountability in the Use of Artificial Intelligence for Public Administrations'. Available at: https://doi.org/10.1145/3461702.3462631.

Lüthi, N., Matt, C. and Myrach, T. (2021) 'A value-sensitive design approach to minimize value tensions in software-based risk-assessment instruments', *Journal of Decision Systems*, 30(2–3), pp. 194–214. Available at: https://doi.org/10.1080/12460125.2020.1859744.

Mackley, A. *et al.* (2023) 'An introduction to social security in the UK'. Available at: https://commonslibrary.parliament.uk/research-briefings/cbp-9535/ (Accessed: 14 August 2023).

Mangal, P., Rajesh, A. and Misra, R. (2020) 'Big Data in Climate Change Research: opportunities and Challenges', in *2020 International Conference on Intelligent Engineering and Management (ICIEM)*. *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, pp. 321–326. Available at: https://doi.org/10.1109/ICIEM48762.2020.9160174.

Marsh, S. and McIntyre, N. (2020) *Nearly half of councils in Great Britain use algorithms to help make claims decisions | Local government | The Guardian*, *The Guardian*. Available at: https://www.theguardian.com/society/2020/oct/28/nearly-half-of-councils-in-great-britain-use-algorithms-to-help-make-claims-decisions (Accessed: 5 November 2020).

Mason, J. (2002) *Qualitative Researching 2nd Edition*.

Mckinsey & Comapny (2022) *Data ethics: What it means and what it takes | McKinsey*. Available at: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/data-ethics-what-it-means-and-what-it-takes (Accessed: 9 March 2023).

Metcalf, J., Moss, E., *et al.* (2021) 'Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York, NY, USA: Association for Computing Machinery (FAccT '21), pp. 735–746. Available at: https://doi.org/10.1145/3442188.3445935.

Metcalf, J., Anne Watkins, E., *et al.* (2021) 'Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts CCS CONCEPTS'. Available at: https://doi.org/10.1145/3442188.3445935.

Monaghan, M. and Ingold, J. (2019) 'Policy practitioners' accounts of evidence-based policy making: The case of universal credit', *Journal of Social Policy*, 48(2), pp. 351–368. Available at: https://doi.org/10.1017/S004727941800051X.

Morgan, D.L. (1998) *The focus group guide book [electronic resource].*, *SAGE*.

Moss, E. *et al.* (2021) *Assembling Accountability: Algorithmic Impact Assessment for the Public Interest*. Available at: https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/ (Accessed: 2 August 2023).

Moss, E. and Metcalf, J. (2020) 'Ethics Owners: A New Model of Organizational Responsibility in Data-Driven Technology Companies'.

Murphy, P. (2001) *Affiliation Bias and Expert Disagreement in Framing the Nicotine Addiction Debate*.

Neville, K.J. and Weinthal, E. (2016) 'Mitigating Mistrust? Participation and Expertise in Hydraulic Fracturing Governance', *Review of Policy Research*, 33(6), pp. 578–602. Available at: https://doi.org/10.1111/ropr.12201.

Newquist, H. (2020) *The Brain Makers: The History of Artifical Intelligence*. 2nd edition. The Relayer Group.

Noble, S.U. (2018) *Algorithms of oppression : how search engines reinforce racism*. Available at: https://www.amazon.co.uk/dp/B075XS7Y7D/ref=dp-kindle-redirect?_encoding=UTF8&btkr=1 (Accessed: 29 March 2019).

Office for Artificial Intelligence (2021) *National AI Strategy*. Available at: https://www.gov.uk/government/publications/national-ai-strategy (Accessed: 25 October 2021).

O'Hara, K. (2019) 'Data Trusts Ethics, Architecture and Governance for Trustworthy Data Stewardship'.

O'Neil, C. (2017) *Weapons of Math Destruction*. London, England: Penguin.

Orr, W. and Davis, J.L. (2020) 'Attributions of ethical responsibility by Artificial Intelligence practitioners', *Information Communication and Society*, 23(5), pp. 719–735. Available at: https://doi.org/10.1080/1369118X.2020.1713842.

Oswald, M. *et al.* (2018) 'Algorithmic risk assessment policing models: lessons from the Durham HART model and "Experimental" proportionality', *Information & Communications Technology Law*, 27(2), pp. 223–250. Available at: https://doi.org/10.1080/13600834.2018.1458455.

Oxford Insights (2019) *Government AI Readiness Index 2019 — Oxford Insights — Oxford Insights*, *Oxford Insights*. Available at: https://www.oxfordinsights.com/ai-readiness2019 (Accessed: 13 November 2020).

Park, S. and Humphry, J. (2019) 'Exclusion by design: intersections of social, digital and data exclusion', *Information Communication and Society*, 22(7), pp. 934–953. Available at: https://doi.org/10.1080/1369118X.2019.1606266.

Peña Gangadharan, S. and Niklas, J. (2019) 'Decentering technology in discourse on discrimination', *Information, Communication & Society*, 22(7), pp. 882–899. Available at: https://doi.org/10.1080/1369118X.2019.1593484.

Pessach, D. and Shmueli, E. (2023) 'A Review on Fairness in Machine Learning', *ACM Computing Surveys*, 55(3), pp. 1–44. Available at: https://doi.org/10.1145/3494672.

Pietsch, W. (2016) 'The Causal Nature of Modeling with Big Data', *Philosophy & Technology*, 29(2), pp. 137–171. Available at: https://doi.org/10.1007/s13347-015-0202-2.

PricewaterhouseCoopers (2017) *PwC's Global Artificial Intelligence Study: Sizing the prize*, *PwC*. Available at: https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html (Accessed: 1 February 2023).

PricewaterhouseCoopers (2019) *Responsible AI Toolkit: PwC*. Available at: https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html (Accessed: 10 December 2019).

Public Law Project (2022) 'Machine learning used to stop Universal Credit payments', *Public Law Project*, 11 July. Available at: https://publiclawproject.org.uk/latest/dwp-accounts-reveal-algorithm-used-to-stop-universal-credit-payments/ (Accessed: 11 August 2023).

Redden, J. and Brand, J. (2017) 'Data harm record', in. Available at: https://www.semanticscholar.org/paper/Data-harm-record-Redden-Brand/91eaa4984520084e62b4a876a68ac23552413da9 (Accessed: 30 August 2023).

Reisman, D. *et al.* (2018) 'Algorithmic impact assessments: A practical framework for public agency accountability', *AI Now Institute*, (April), p. 22.

Rorie, M. and West, M. (2022) 'Can "Focused Deterrence" Produce More Effective Ethics Codes? An Experimental Study', *Journal of White Collar and Corporate Crime*, 3(1), pp. 33–45. Available at: https://doi.org/10.1177/2631309X20940664.

Rovatsos, D.M., Mittelstadt, D.B. and Koene, D.A. (2019) 'Landscape Summary : Bias in Algorithmic'.

Schuler, D. and Namioka, A. (eds) (1993) *Participatory design : Principles and practices*.

Scott, R. and Davies, G. (2006) *Organizations : rational, natural, and open systems*. Englewood Cliffs, N.J.: Prentice Hall.

Seaver, N. (2017) 'Algorithms as culture: Some tactics for the ethnography of algorithmic systems', *Big Data & Society*, 4(2), p. 205395171773810. Available at: https://doi.org/10.1177/2053951717738104.

Selbst, A.D. *et al.* (2018) 'Fairness and Abstraction in Sociotechnical Systems', in *2019 ACM Conference on Fairness, Accountability, and Transparency (FAT*)*. Available at: https://papers.ssrn.com/abstract=3265913 (Accessed: 11 November 2020).

Sloane, M. (2020) *Participation-washing could be the next dangerous fad in machine learning*, *MIT Technology Review*. Available at: https://www.technologyreview.com/2020/08/25/1007589/participation-washing-ai-trends-opinion-machine-learning/ (Accessed: 8 May 2023).

Sloane, M. *et al.* (2022) 'Participation Is not a Design Fix for Machine Learning', in *Equity and Access in Algorithms, Mechanisms, and Optimization*. *EAAMO '22: Equity and Access in*

Algorithms, Mechanisms, and Optimization, Arlington VA USA: ACM, pp. 1–6. Available at: https://doi.org/10.1145/3551624.3555285.

Star, S.L. (1989) 'Chapter 2 - The Structure of Ill-Structured Solutions: Boundary Objects and Heterogeneous Distributed Problem Solving', in L. Gasser and M.N. Huhns (eds) *Distributed Artificial Intelligence*. San Francisco (CA): Morgan Kaufmann, pp. 37–54. Available at: https://doi.org/10.1016/B978-1-55860-092-8.50006-X.

Steedman, R., Kennedy, H. and Jones, R. (2020) 'Complex ecologies of trust in data practices and data-driven systems', *Information, Communication & Society*, 23(6), pp. 817–832. Available at: https://doi.org/10.1080/1369118X.2020.1748090.

The Equality Act (2010) *Equality Act 2010*. Statute Law Database. Available at: https://www.legislation.gov.uk/ukpga/2010/15/contents (Accessed: 22 August 2023).

Treasury Board of Canada Secretariat (2021) *Algorithmic Impact Assessment Tool*. Available at: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html (Accessed: 3 February 2023).

Trendall, S. (2019) *DWP digital chief picks AI, data sharing, and digital identity as 2019 priorities | PublicTechnology.net*, *Public Technology*. Available at: https://www.publictechnology.net/articles/news/dwp-digital-chief-picks-ai-data-sharing-and-digital-identity-2019-priorities (Accessed: 29 October 2019).

UK Government (2018) 'Centre for Data Ethics and Innovation (CDEI) - GOV.UK', (March). Available at: https://www.gov.uk/government/groups/centre-for-data-ethics-and-innovation-cdei.

Umbrello, S. and van de Poel, I. (2021) 'Mapping value sensitive design onto AI for social good principles', *AI and Ethics*, 1(3), pp. 283–296. Available at: https://doi.org/10.1007/s43681-021-00038-3.

Van Audenhove, L. and Donders, K. (2019) 'Talking to People III: Expert Interviews and Elite Interviews', in *The Palgrave Handbook of Methods for Media Policy Research*. Springer International Publishing, pp. 179–197. Available at: https://doi.org/10.1007/978-3-030-16065-4_10.

Veale, M. and Binns, R. (2017) 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data', *Big Data & Society*, 4(2), p. 205395171774353. Available at: https://doi.org/10.1177/2053951717743530.

Veale, M., Van Kleek, M. and Binns, R. (2018) 'Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making', pp. 1–14. Available at: https://doi.org/10.1145/3173574.3174014.

Verbeek, S. and Groeneveld, S. (2012) 'Do "hard" diversity policies increase ethnic minority representation?: An assessment of their (in)effectiveness using administrative data',

Personnel Review, 41(5), pp. 647–664. Available at: https://doi.org/10.1108/00483481211249157.

Vogel, D. (2006) *The Market for Virtue: The Potential and Limits of Corporate Social Responsibility*, *Brookings Insitituion Press*. Brookings Institution Press. Available at: https://www.jstor.org/stable/10.7864/j.ctt6wpg2c (Accessed: 27 March 2021).

Wachter, S., Mittelstadt, B. and Russell, C. (2017) 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR', *SSRN Electronic Journal*, pp. 1–52. Available at: https://doi.org/10.2139/ssrn.3063289.

Waterfield, S. (2023) 'DWP ramps up investment in AI models to detect fraud despite algorithmic bias concerns', *Tech Monitor*, 8 July. Available at: https://techmonitor.ai/leadership/digital-transformation/dwp-ai-fraud-bias (Accessed: 11 August 2023).

Wing, J. (2018) *Data for Good: FATES, Elaborated*, *The Data Science Institute at Columbia University*. Available at: https://datascience.columbia.edu/news/2018/data-for-good-fates-elaborated/ (Accessed: 9 March 2023).

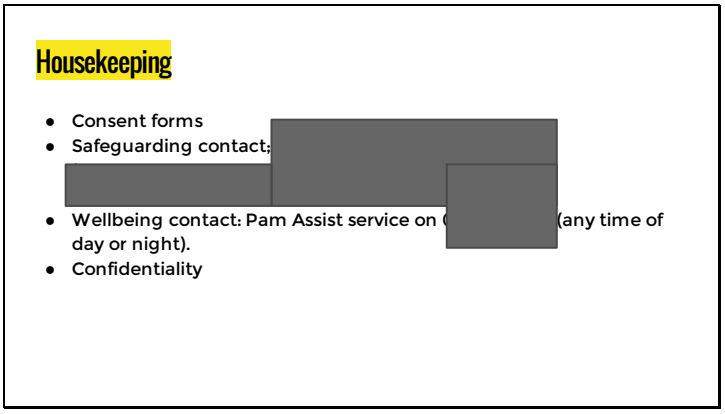
Yam, J. and Skorburg, J.A. (2021) 'From human resources to human rights: Impact assessments for hiring algorithms', *Ethics and Information Technology*, 23(4), pp. 611–623. Available at: https://doi.org/10.1007/s10676-021-09599-7.

Appendix: A – Paper three educational workshop slides

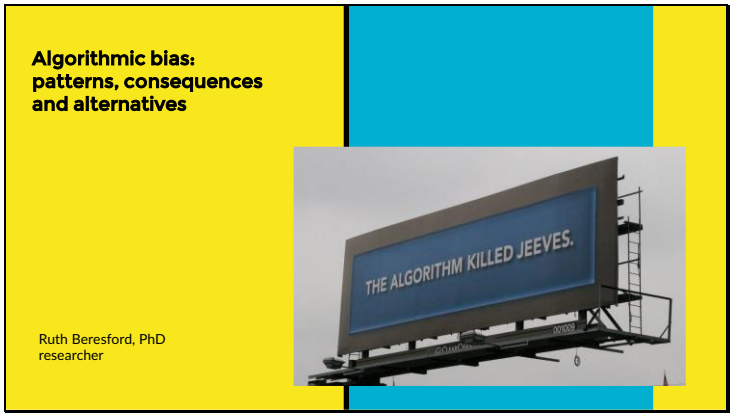
Workshop one (WS1)

Appendix

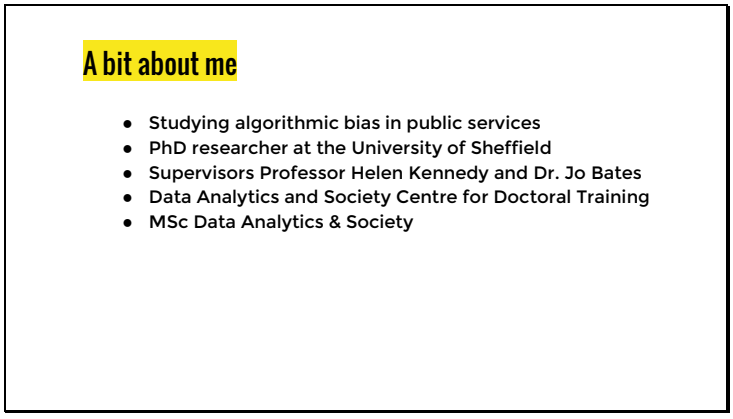
Slide 1



Slide 2



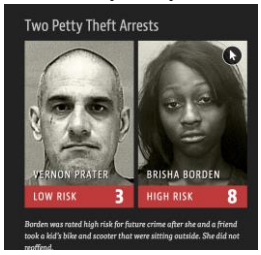
Slide 3



Slide 4

Examples of Algorithmic Bias in the public sector

Criminal justice system



A-level algorithm

Slide 5

What is an algorithm? What is algorithmic bias?

- An algorithm is a "description of the method by which a task is to be accomplished" (Goffey, 2006)
- "[A] senior software engineer with a prestigious undergraduate degree in computer science told me that her training on algorithms in theory was irrelevant to her work on algorithms in practice, because **algorithms in practice were harder to precisely locate** [..] The 'algorithm' here was **a collective product, and consequently every- one felt like an outsider to it"** (Seaver, 2017)
- "[p]ublic discussions of algorithmic bias currently **conflate many different types, sources, and impacts of biases**, with the net result that **the term has little coherent content**" (Danks and London, 2017, p2)

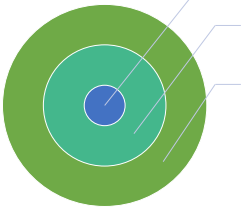
Slide 6

My working definition

- "unfair outcomes caused by the automation of statistical techniques which aim to predict outcomes based on historical data and probabilistic understandings" (Beresford, 2020)

Slide 7

Where does algorithmic bias come from?



Algorithm/model frame: e.g. binary v non-binary classification, bias towards popularity (i.e. recommendation/search algorithms)

Data frame: e.g. biased data, historical data, poor use of proxy data

Socio-technical frame: e.g. reproducing historical social inequalities, biases of data scientists, biases of those acting on results

(J. Bates, 2018, adapted from Selbst, Boyd, Friedler, Venkatasubramanian, & Janet, 2018)

Slide 8

Algorithmic Bias PhD

Paper	Research Questions	Partner Output
Paper 1 (DWP)	RQ1: What aspects of frameworks for mitigating algorithmic bias are most relevant to DWP? RQ2: Within these aspects/frameworks, what are the limitations to using a technical lens on bias?	Report on Bias
Paper 2 (Aurora AI & Algorithmic Justice Orgs)	RQ1: How do public services using data-driven technologies attempt to mitigate algorithmic bias? RQ2: Which attempts to mitigate algorithmic bias might be considered successful, and why?	Report on a case study of the Aurora AI project by the Finnish Ministry of Finance
Paper 3	RQ1: How might successful approaches to algorithmic bias mitigation be incorporated into DWP practice?	A set of practical guidelines for fairer algorithmic systems from a sociotechnical viewpoint

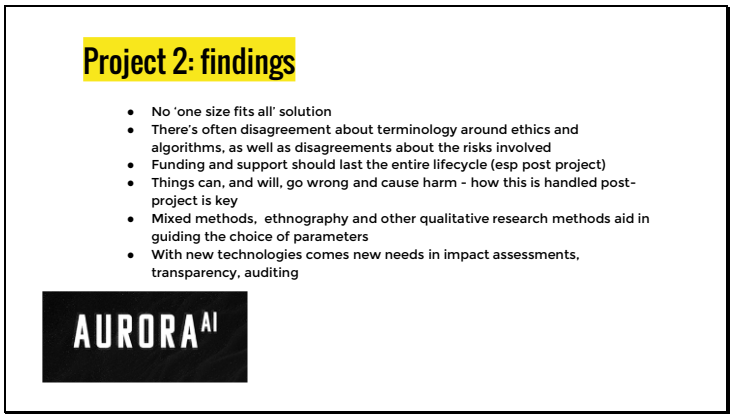
Slide 9

Project 1: findings

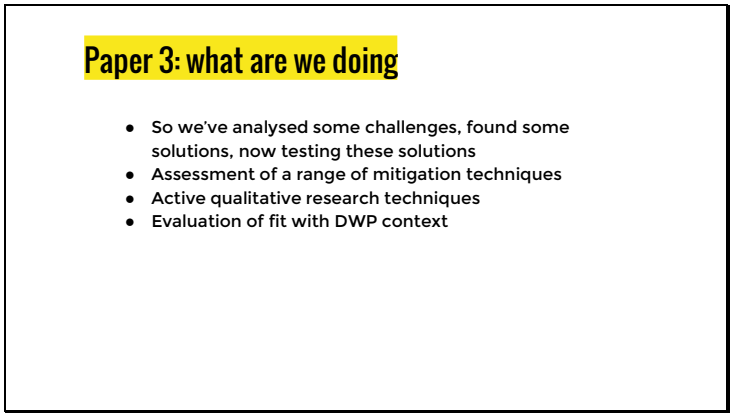
- Social and technological solutions are context sensitive
- Each environment each project is created in is different
- Data governance frameworks are incredibly useful for staff
- Stakeholder engagement is important
- Outside organisations can provide external insight, and be used to increase understanding of stakeholders needs

Department for Work & Pensions

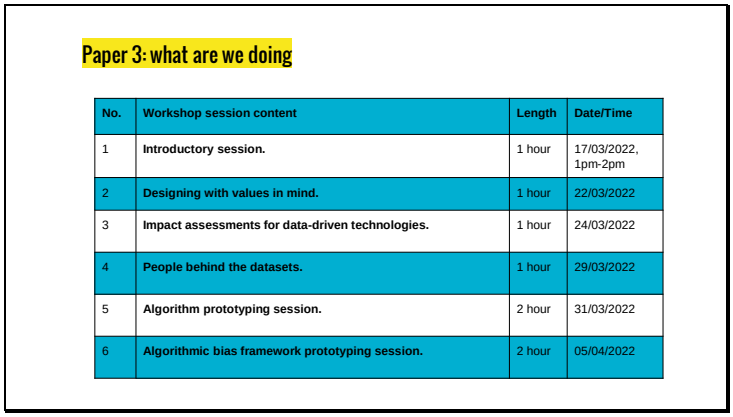
Slide 10



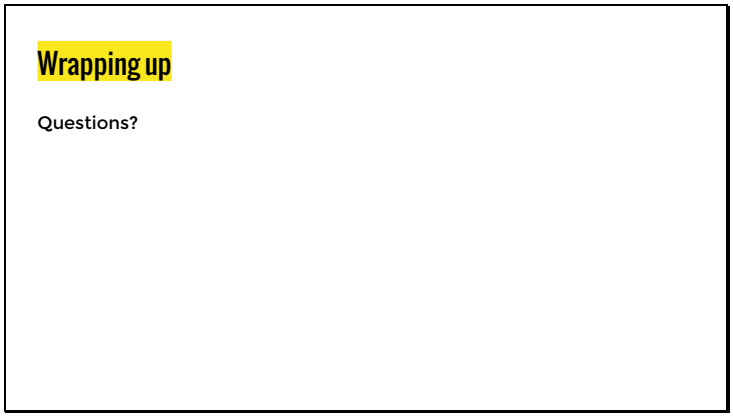
Slide 11



Slide 12



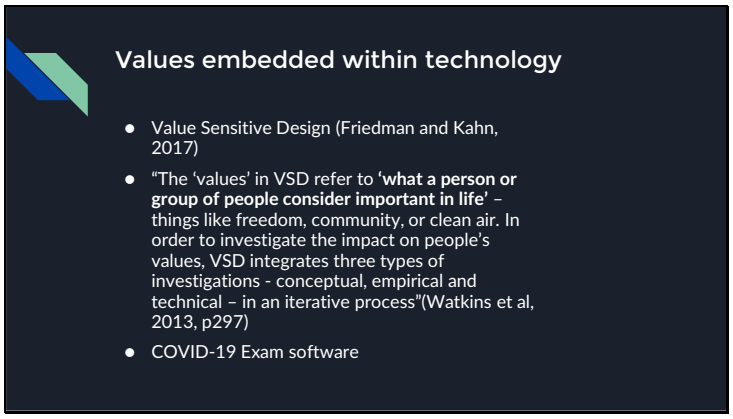
Slide 13



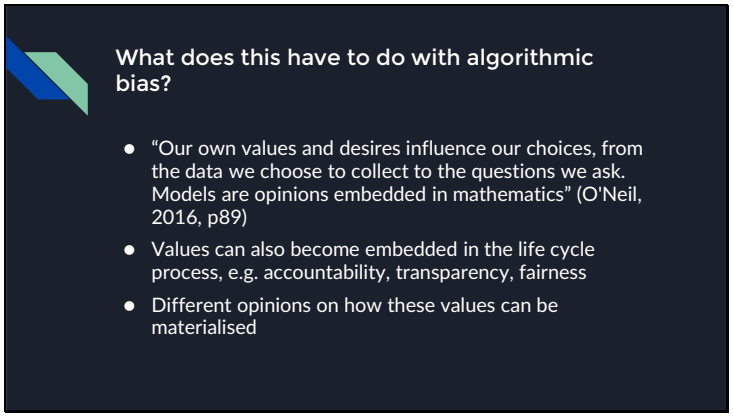
Workshop two (WS2)

Slide 1

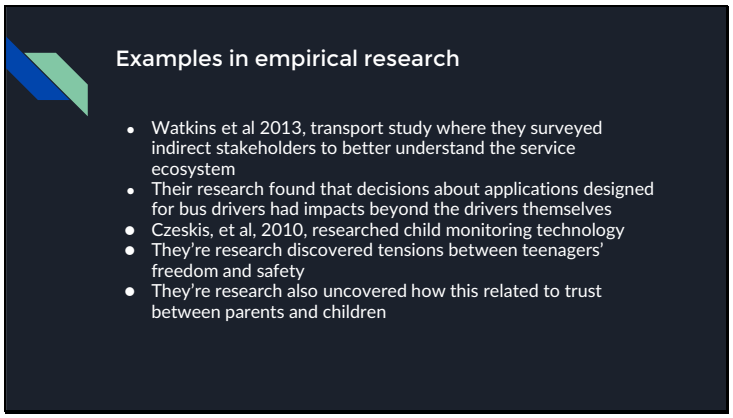
Slide 2



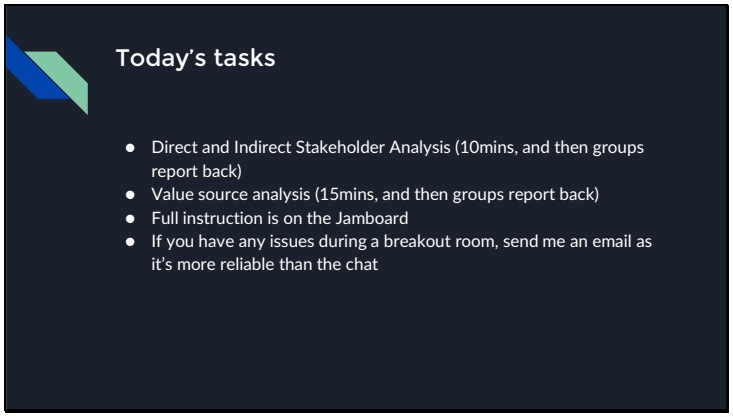
Slide 3



Slide 4



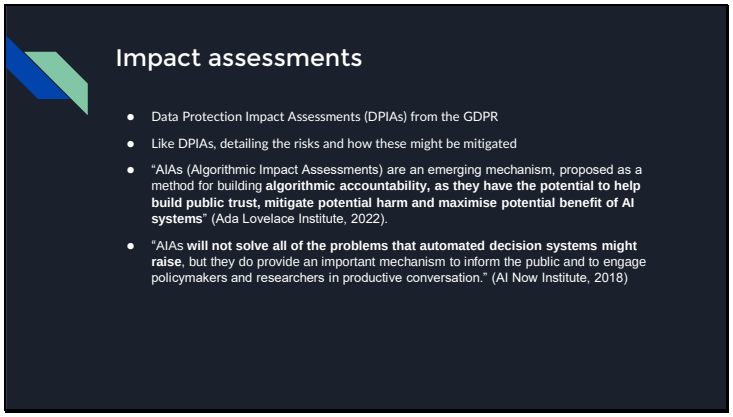
Slide 5



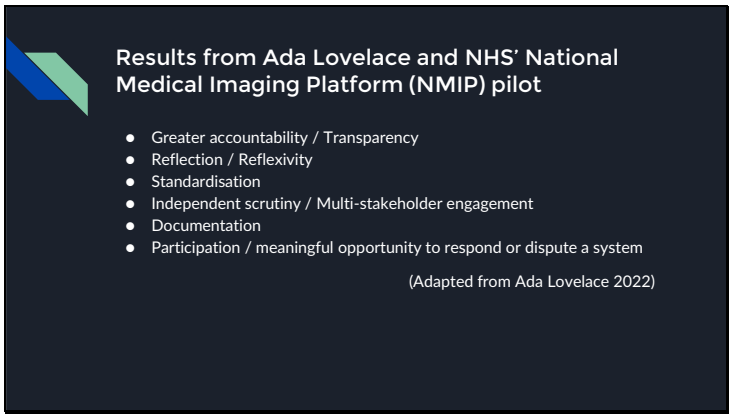
Workshop three (WS3)

Slide 1

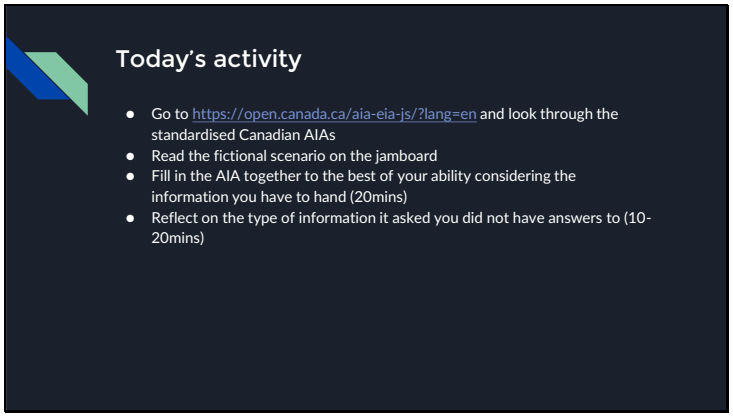
Slide 2



Slide 3



Slide 4

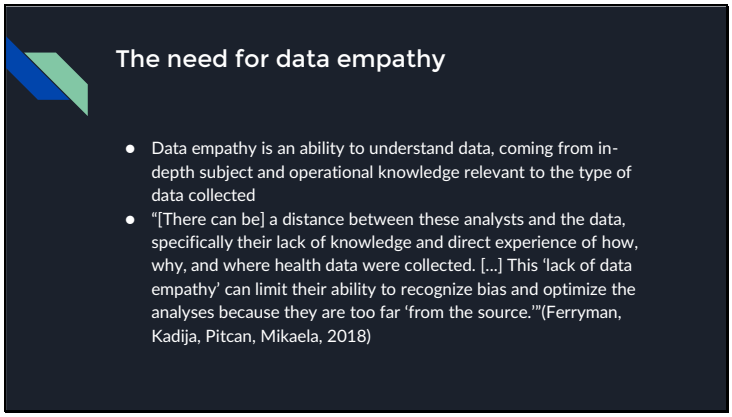


Workshop four (WS4)

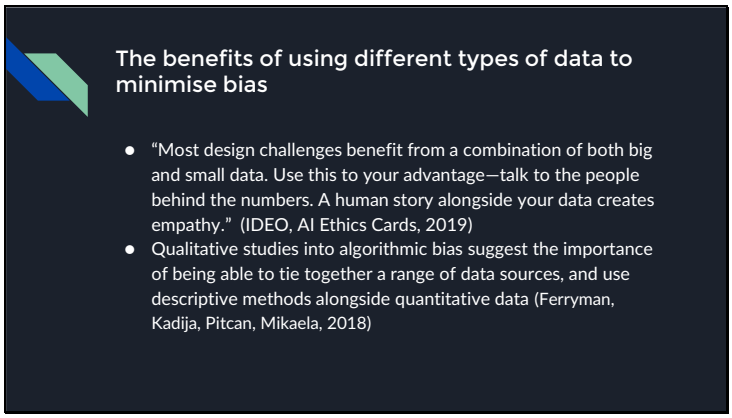
Slide 1

Slide 2

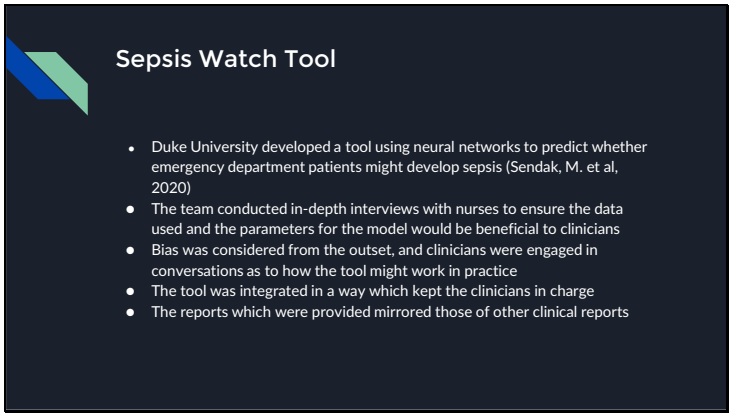
Slide 3



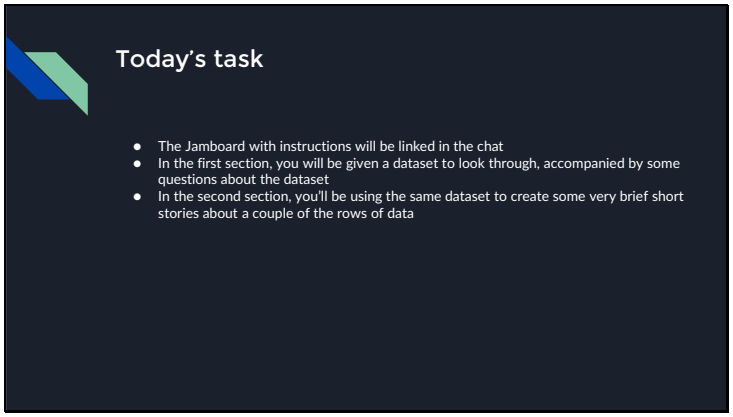
Slide 4



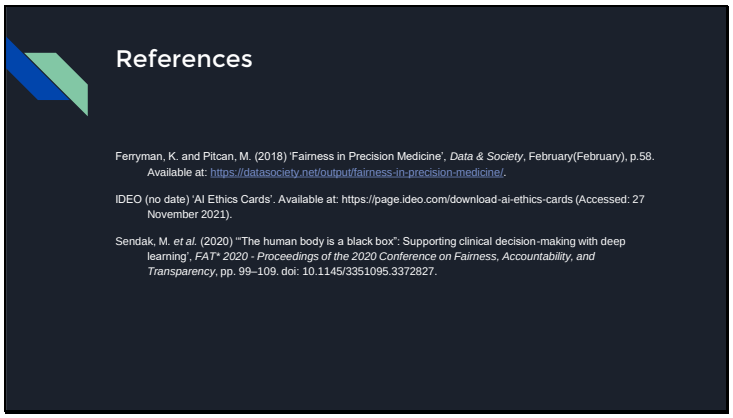
Slide 5



Slide 6



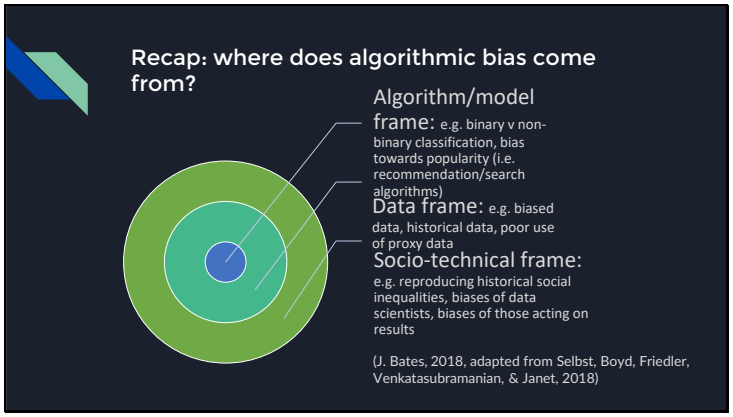
Slide 7



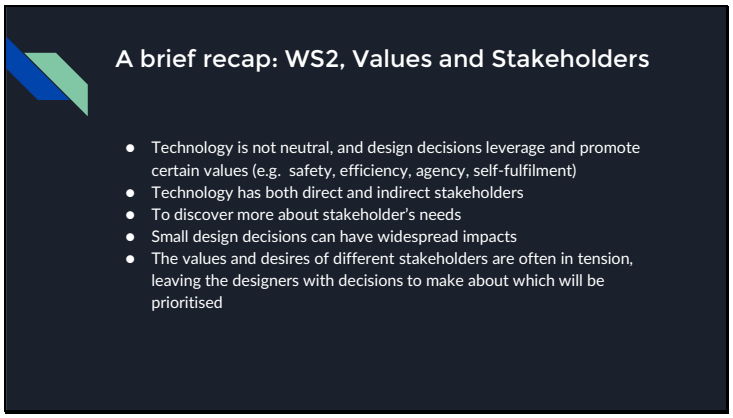
Workshop five (WS5)

Slide 1

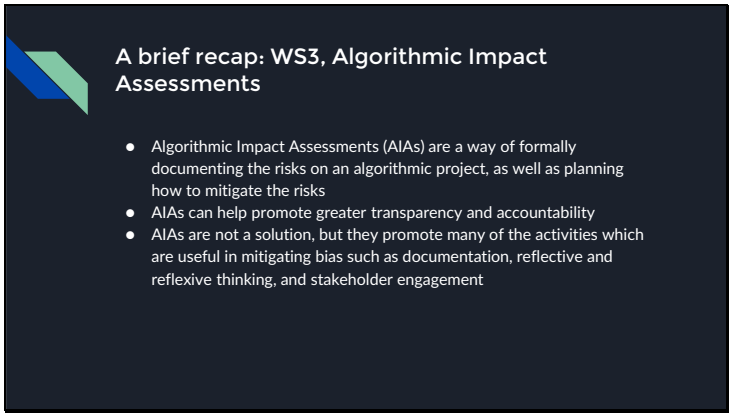
Slide 2



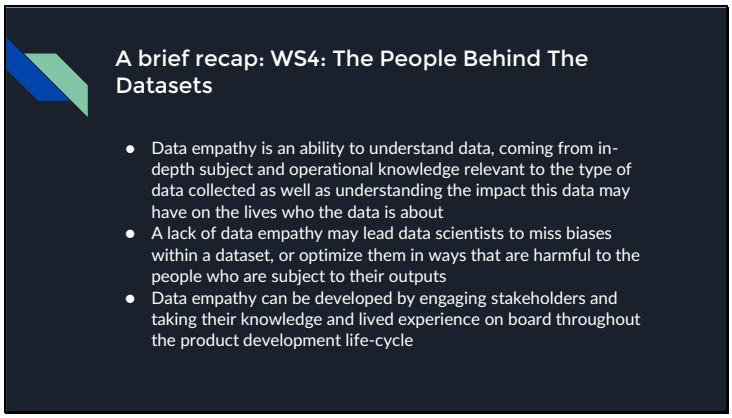
Slide 3



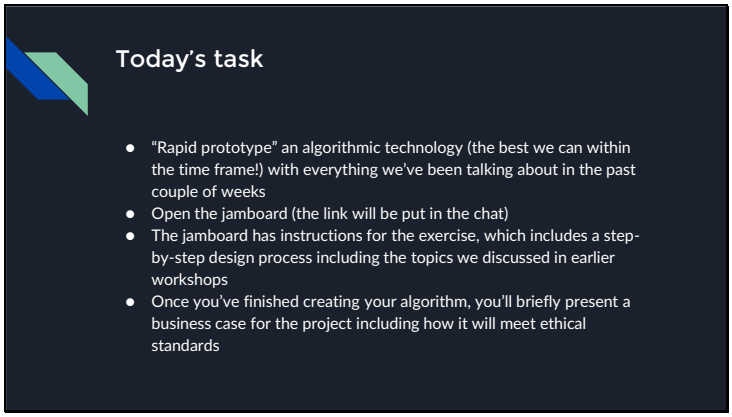
Slide 4



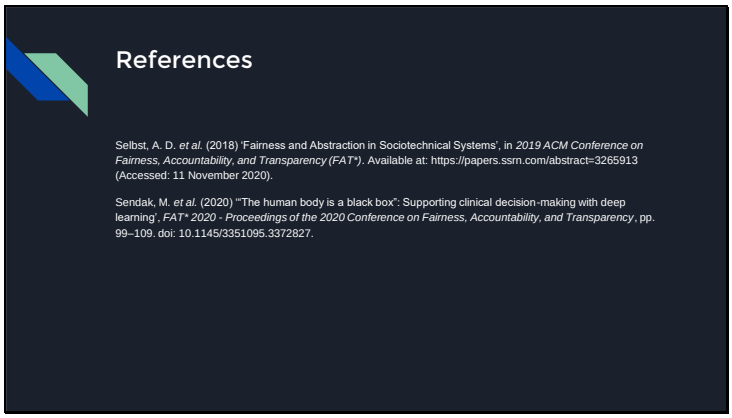
Slide 5



Slide 6



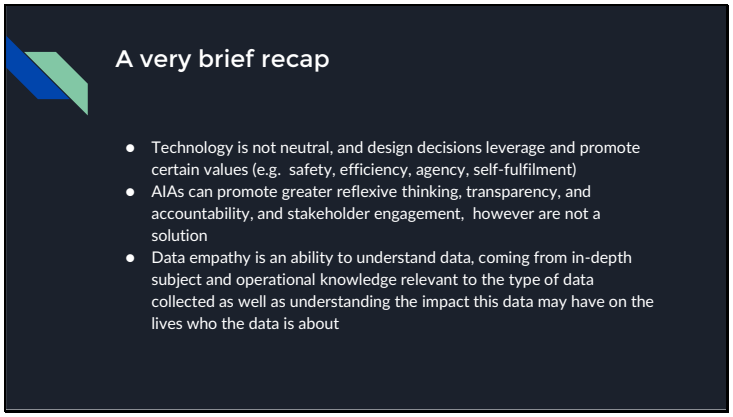
Slide 7



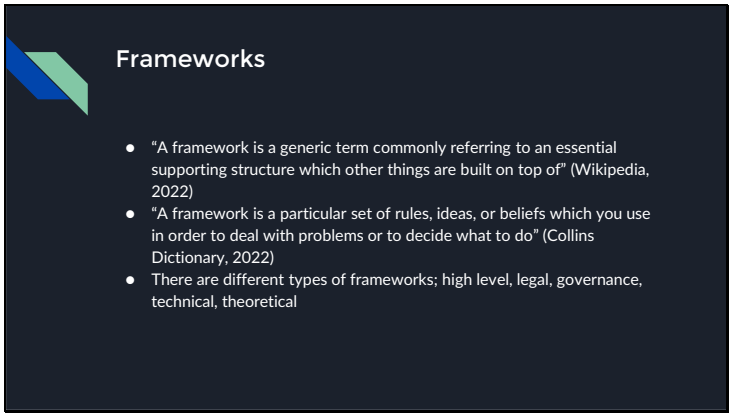
Workshop six (WS6)

Slide 1

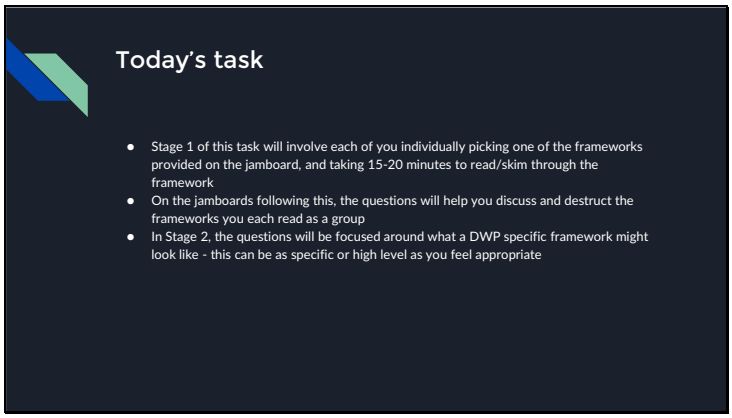
Slide 2



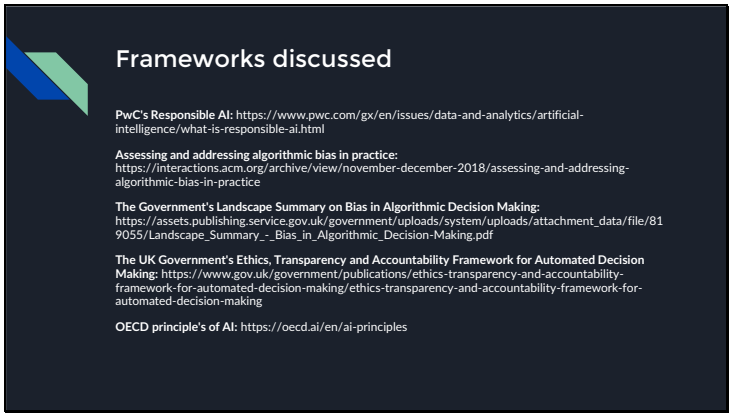
Slide 3



Slide 4

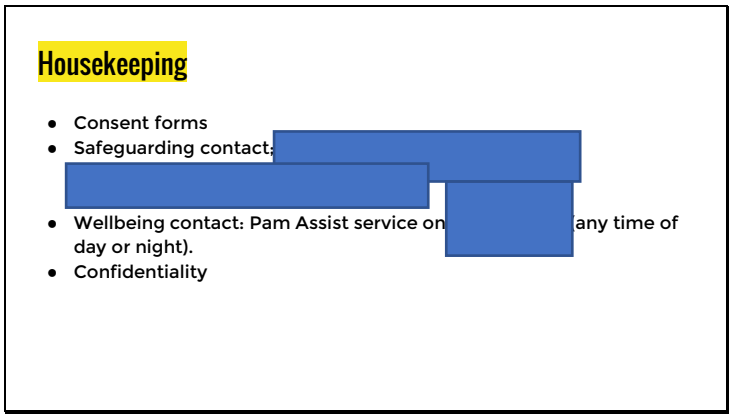


Slide 5

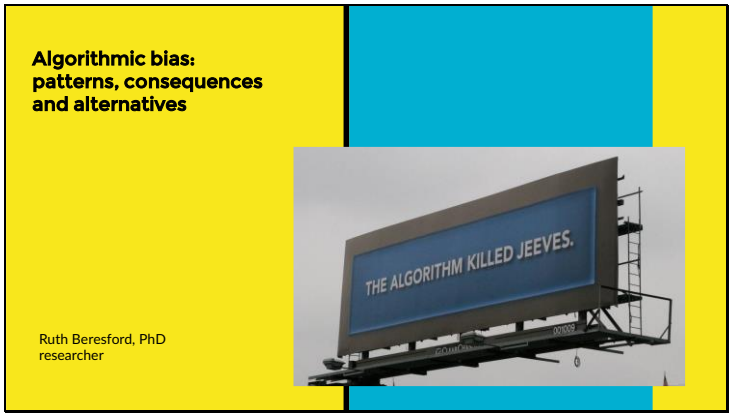


Workshop seven (WS7)

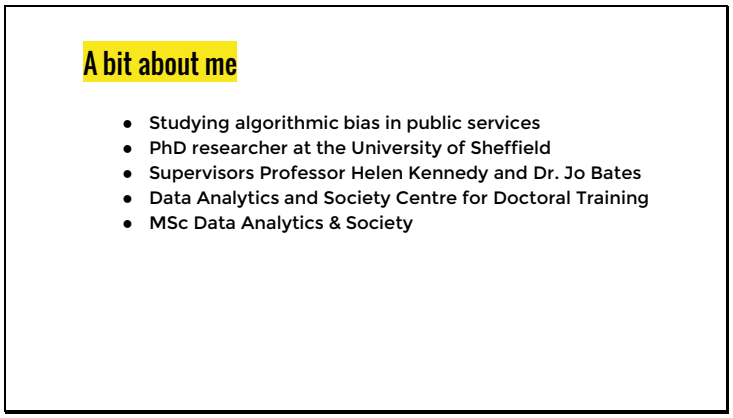
Slide 1



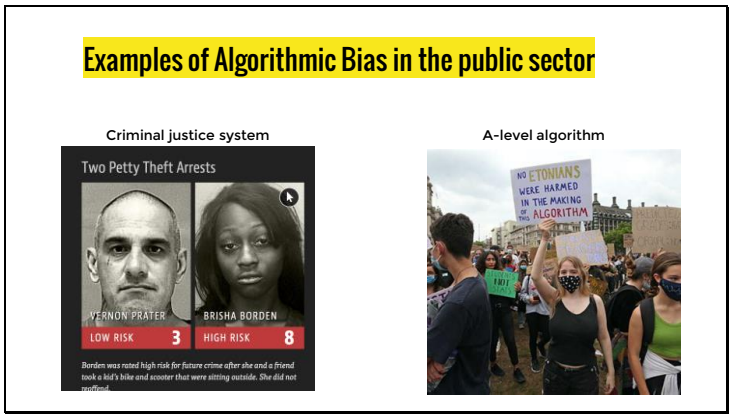
Slide 2



Slide 3



Slide 4



Slide 5

What is an algorithm? What is algorithmic bias?

- An algorithm is a "description of the method by which a task is to be accomplished" (Goffey, 2006)
- "[A] senior software engineer with a prestigious undergraduate degree in computer science told me that her training on algorithms in theory was irrelevant to her work on algorithms in practice, because **algorithms in practice were harder to precisely locate** [..] The 'algorithm' here was **a collective product, and consequently every- one felt like an outsider to it"** (Seaver, 2017)
- "[p]ublic discussions of algorithmic bias currently **conflate many different types, sources, and impacts of biases**, with the net result that **the term has little coherent content**" (Danks and London, 2017, p2)

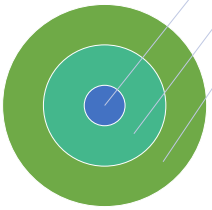
Slide 6

My working definition

- "unfair outcomes caused by the automation (?) of statistical techniques which aim to predict outcomes based on historical data and probabilistic understandings" (Beresford, 2020)

Slide 7

Where does algorithmic bias come from?



Algorithm/model frame: e.g. binary v non-binary classification, bias towards popularity (i.e. recommendation/search algorithms)

Data frame: e.g. biased data, historical data, poor use of proxy data

Socio-technical frame: e.g. reproducing historical social inequalities, biases of data scientists, biases of those acting on results

(J. Bates, 2018, adapted from Selbst, Boyd, Friedler, Venkatasubramanian, & Janet, 2018)

Slide 8

Algorithmic Bias PhD

Paper	Research Questions	Partner Output
Paper 1 (DWP)	RQ1: What aspects of frameworks for mitigating algorithmic bias are most relevant to DWP? RQ2: Within these aspects/frameworks, what are the limitations to using a technical lens on bias?	Report on Bias
Paper 2 (Aurora AI & Algorithmic Justice Orgs)	RQ1: How do public services using data-driven technologies attempt to mitigate algorithmic bias? RQ2: Which attempts to mitigate algorithmic bias might be considered successful, and why?	Report on a case study of the Aurora AI project by the Finnish Ministry of Finance
Paper 3	RQ1: How might successful approaches to algorithmic bias mitigation be incorporated into DWP practice?	A set of practical guidelines for fairer algorithmic systems from a sociotechnical viewpoint

Slide 9

Project 1: findings

- Data science project at DWP
- Solutions are context sensitive
- Data governance frameworks are incredibly useful for staff
- Stakeholder engagement is important
- Outside organisations can provide external insight, and be used to increase understanding of stakeholders needs

Department for Work & Pensions

Slide 10

Project 2: findings

- No 'one size fits all' solution
- There's often disagreement about, as well as disagreements about the risks involved
- Funding and support should last the entire lifecycle (esp post project)
- Things can, and will, go wrong and cause harm - how this is handled post-project is key
- Mixed methods, ethnography and other qualitative research methods aid in guiding the choice of parameters
- Value sensitive design, algorithmic impact assessments, data empathy

AURORA AI

Slide 11

Paper 3: what are we doing

- So we've analysed some challenges, found some solutions, now testing these solutions
- Assessment of a range of mitigation techniques
- Active qualitative research techniques
- Evaluation of fit with DWP context
- Follow up interviews

Slide 12

Paper 3: what are we doing

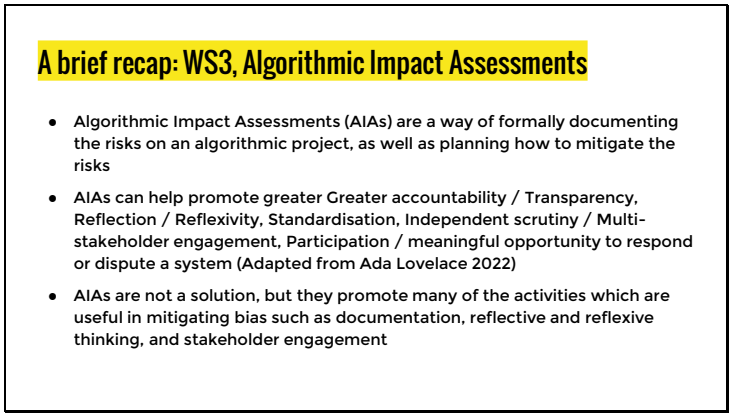
No.	Workshop session content	Length	Date/Time
1	Introductory session.	1 hour	17/03/2022, 1pm-2pm
2	Designing with values in mind.	1 hour	22/03/2022
3	Impact assessments for data-driven technologies.	1 hour	24/03/2022
4	People behind the datasets and data empathy	1 hour	29/03/2022
5	Algorithm prototyping session.	2 hour	31/03/2022
6	Algorithmic bias framework prototyping session.	2 hour	05/04/2022

Slide 13

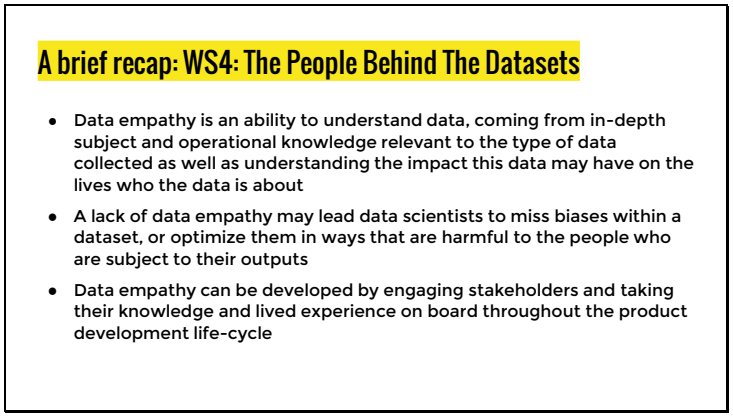
A brief recap: WS2, Values and Stakeholders

- Technology is not neutral, and design decisions leverage and promote certain values (e.g. safety, efficiency, agency, self-fulfilment)
- Technology has both direct and indirect stakeholders
- Small design decisions can have widespread impacts
- The values and desires of different stakeholders are often in tension, leaving the designers with decisions to make about which will be prioritised

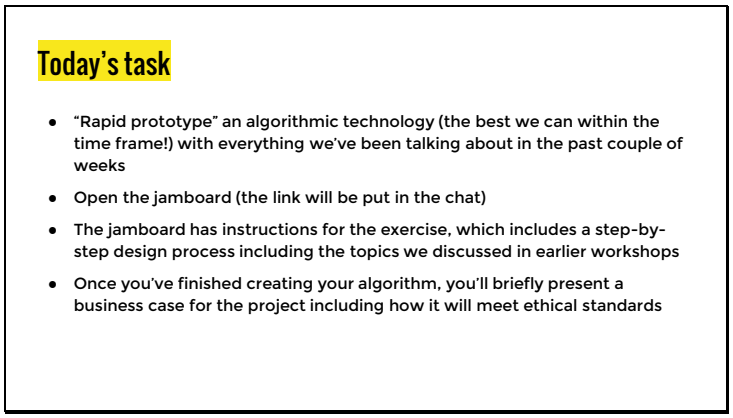
Slide 14



Slide 15

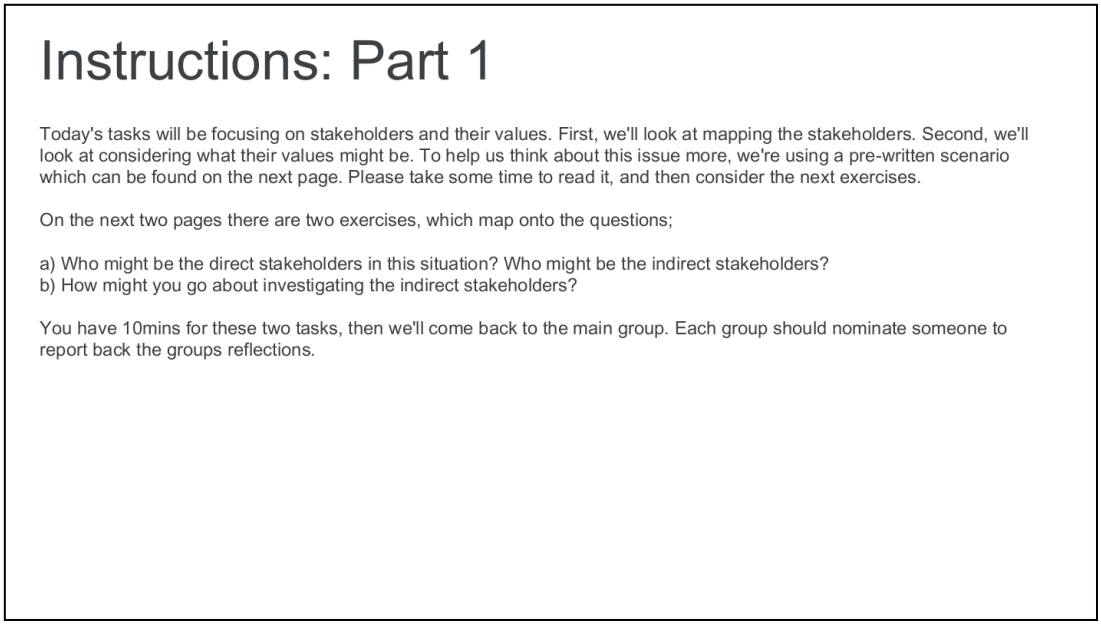


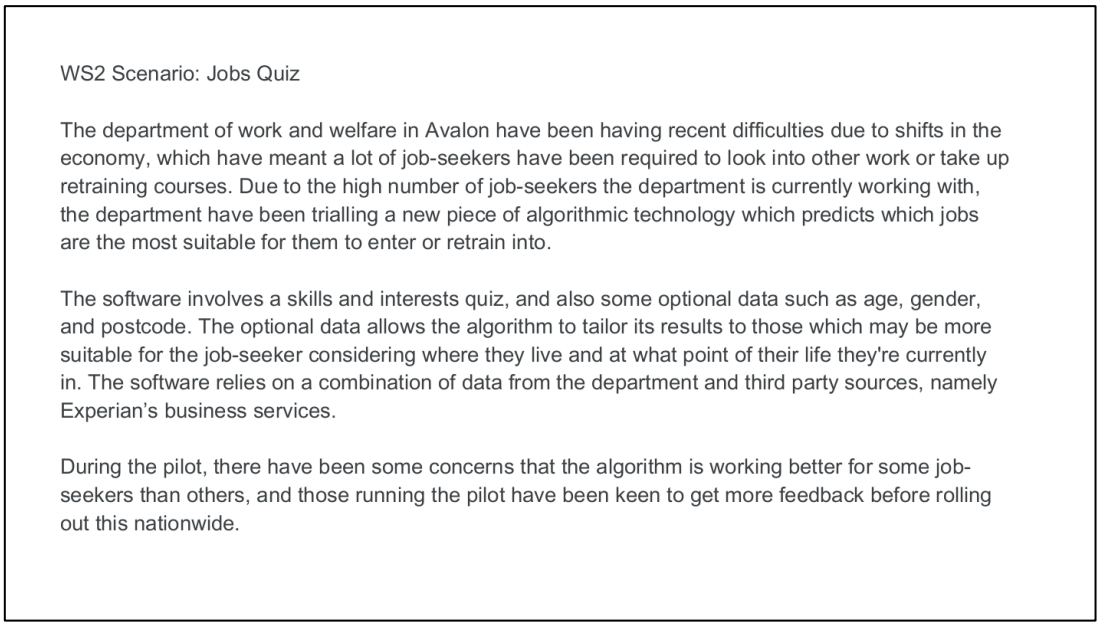
Slide 16



Appendix: B – Paper three, educational workshop activity jamboards

Workshop two (WS2)





Stakeholder statements from the pilot:

Client 1#
I really liked the quiz, think it helped me think more about the skills I've picked up in my past jobs. I used to work in a call centre for an energy supplier, but they downsized in the pandemic and it seems safer to take voluntary redundancy.

The job quiz has given me more confidence when putting my cv together. I have got 2 interviews since I widened my job search. Haven't heard back from either of them, which I'm a bit worried about with my mental health, might still take a while to land on my feet, but hopefully ill get there soon.

Client 2#
Some of the questions felt intrusive. I've been looking after our children for the past few years, and am currently looking for work as my partner got made redundant. The algorithm suggested a lot of care work options, despite having a degree in illustration from 10 years ago. I've kept up my art work on the side, and I was expecting more help trying to get into an industry which is more suitable to my skillset.

Work coach:
The new system has definitely reduced some of our workloads, and I'm happy to see it's really working for some customers. There's a few issues I'm hoping they'll get around to fixing though I've mentioned them to my manager but haven't heard much back about them. The system isn't as localised as we'd been told it was going to be, so sometimes we get customers getting suggestions that aren't really viable in this area - like fruit picking. I've been a bit concerned it's also killed the confidence of some of my customers, I've had to stress it's more of an aid than predictive.

Who might be the direct stakeholders in this situation? Who might be the indirect stakeholders?
Direct stakeholders: People, organisations or groups who are either directly involved in the design, or directly impacted by the technology

Indirect stakeholders: People, organisations or groups who may not be directly involved or impacted, but who's work or day-to-day life may be influenced in some way

Finding out who the indirect stakeholders of a project are trickier than the direct stakeholders, however these are still people and organisations which might be impacted by this technology. What do you think the design team in the scenario could do to uncover more of these indirect stakeholders? How could they discover more about their needs? Jot down some ideas on post-it notes below.

Instructions: Part 2

So, now we've got our direct and indirect stakeholders sorted, it's time to think about values! On the next two pages there's two questions :

– During the design process of this algorithmic technology, what values does it seem to support?
– Considering both your direct and indirect stakeholders, what values might be in tension between these stakeholders?

Remember from the short talk, values can be anything which mean something to someone or an organisation. If you feel like it'll be important to someone in this situation, pop it on the board.

You have 15mins for these two tasks, then we'll come back to the main group. Each group should nominate someone to report back the groups reflections.

From reading the scenario, what values does it sound like the organisation is trying to promote? What might the values be of the designers who originally came up with the project? Take a moment to think about this as a group, and put your ideas down using the post-it notes below.

Considering both your direct and indirect stakeholders, what values might be in tension between any of these stakeholders? Do they all want the same thing? Might their needs and expectations clash? Jot down your ideas using the post-it notes below.

Workshop three (WS3)

Instructions

So, in this workshop we'll be reflecting on an algorithmic impact assessment. To complete this exercise, you'll need to go to the link here: https://open.canada.ca/aia-eia-js/?lang=en. This is the Canadian government's online assessment tool and includes questions about the project the designers have in mind, to assess what steps need to be put into practice to meet the Canadian government's standards. While this is specific to Canada's frameworks, it will still help us think through how these assessments work.

Now, on the next page, there is a small scenario about a fictional algorithm being designed. Please read through this. Once you've read through this, open up the government of Canada's algorithmic impact assessment questionnaire tool. Once you've opened this, as a group, try and fill in the questionnaire using the scenario material the best you can. In some places, you may not feel you can answer some of the questions - make notes of these on the slide following the scenario. This should take about 20mins. If you feel uncomfortable putting answers directly into the tool, please feel free to discuss them and use the jamboard to make notes instead. We might not get through the whole assessment today, however it will provide a taste of an active and standardised AIA (algorithmic impact assessment). Similarly, feel free to decide if you want to skip discussing a question which doesn't seem relevant.

Once you've done this, move to slide 4. On slide 4 you'll be reflecting on the list of unknowns you created in the first section of the exercise. How might the designers of this algorithm find the answers to these questions? Who would they need to talk to? Would any new technical, or organisational procedures need to be put in to do so?

Today's scenario

In Avolon, the department of work and welfare have recently uncovered a series of overpayments to people claiming job-seeking benefits. The problem looks to have occurred due to the way some local authorities handled housing data, which has meant some job-seekers were entitled to less financial assistance than they recieved.

All affected citizens need to be contacted to explain the error, however the department wants to ensure proper care and attention is taken to inform citizens who may still be out of work, have health difficulties or other complex needs. While the department usually contacts people by post in these circumstances, they want to contact those individuals by phone or through a job centre if they can.

To help them do this, they start an algorithmically focused project which sorts through 'regular' individuals, and 'needs extra care' individuals. After these individuals are sorted, the system will then automatically send a letter to each of the individuals in the 'regular' process group a letter stating the mistake and the amount now owned to the department, and those in the 'need extra care' group will be contacted by phone or in person.

Take 20mins to complete the Canadian government's online assessment tool. If you'd like to make notes while discussing it, feel free to use the space below to do so.

Using the fictional scenario, which of the questions on the algorithmic impact assessment questionnaire can you answer? Which did you struggle with? Think about these as a group and put your thoughts down using the post-it notes below.

Now you'll be reflecting on the list of unknowns you created in the first section of the exercise. How might the designers of this algorithm find the answers to these questions? Who would they need to talk to? Would any new technical, or organisational procedures or working practices need to be created to do so?

Workshop four (WS4)

Instructions

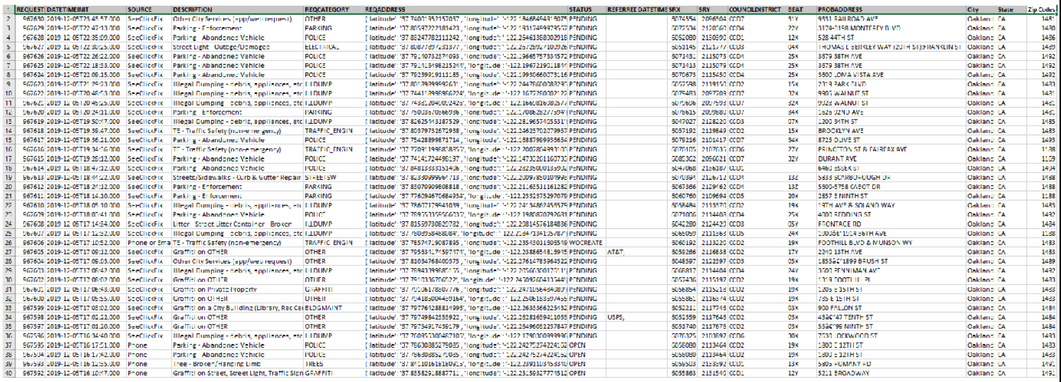
So, today we're looking at developing our data empathy and getting into the people behind the numbers! To do this, we're going to be looking at a datasets, and asking ourselves questions about what we might know about the people who contributed towards the dataset.

On slides 2 and 3, you'll see a dataset relating to an American public service. Look through these columns, and then consider the following questions:

– Where does this dataset come from? Who might be the owner of this dataset? What might the datasets purpose be? (Jamboard slide 4) (5mins)
– For 2 or 3 of the variables, write down some ideas of what these variables might mean. Why might these be logged? Who are these variables useful for? Why are they stored the way they are? (Jamboard slide 5 & 6) (10mins)

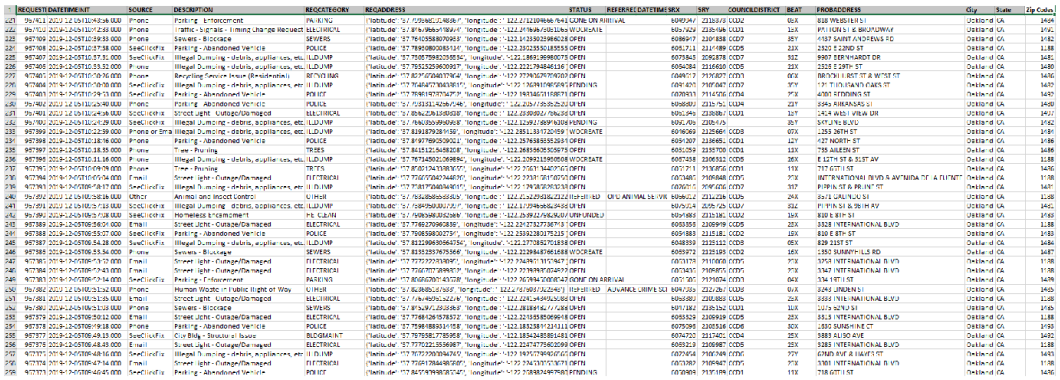
Once you've completed these two exercises, we'll be doing something a little different. For the exercise on Jamboard slide 7, you'll be picking 2 or 3 rows from the dataset and jotting down some ideas of the event the row describes. Consider what the area might be like, what might be happening in the area considering the information you've gathered from the other rows of data, what it might be like to live in this place and what your attitudes might be if you did. Tip, google is a great source of inspiration. (20mins)

American Public Service dataset 1(a):



1	REQUESTID	DATETIMEINIT	SOURCE	DESCRIPTION	REQCATEGORY	REQADDRESS	STATUS	REFERREE DATETIME	SRX	SRY	COUNCILDISTRICT	BEAT	PROBADDRESS	City	State	Zip Codes
2	967630	2019-12-05T23:45:57.000	SeeClickFix	Other City Services (app/web request)	OTHER	{'latitude': '37.74001952157057', 'longitude': '-122.1846849491507	PENDING		6074554	2096504	CCD7	31Y	9931 RAILROAD AVE	Oakland	CA	1481
3	967629	2019-12-05T22:47:13.000	SeeClickFix	Parking - Enforcement	PARKING	{'latitude': '37.80597222181423', 'longitude': '-122.1931749997956	PENDING		6072534	2120060	CCD4	22Y	3174-3198 MONTEREY BLVD	Oakland	CA	1480
4	967628	2019-12-05T22:35:09.000	SeeClickFix	Parking - Abandoned Vehicle	POLICE	{'latitude': '37.88247782111242', 'longitude': '-122.2546138800291	PENDING		6052080	2130890	CCD1	12X	528 44TH ST	Oakland	CA	1486
5	967627	2019-12-05T22:30:25.000	SeeClickFix	Street Light - Outage/Damaged	ELECTRICAL	{'latitude': '37.80877897231377', 'longitude': '-122.2672692710092	PENDING		6051145	2121777	CCD3	04X	THOMAS L BERKLEY WAY (20TH ST):FRANKLIN ST	Oakland	CA	1489
6	967626	2019-12-05T22:26:22.000	SeeClickFix	Parking - Abandoned Vehicle	POLICE	{'latitude': '37.79140732374093', 'longitude': '-122.1966575733457	PENDING		6071431	2115075	CCD4	25X	3579 38TH AVE	Oakland	CA	1492
7	967625	2019-12-05T22:18:33.000	SeeClickFix	Parking - Abandoned Vehicle	POLICE	{'latitude': '37.79141549821524', 'longitude': '-122.1967219011184	PENDING		6071413	2115079	CCD4	25X	3579 38TH AVE	Oakland	CA	1492
8	967624	2019-12-05T22:09:15.000	SeeClickFix	Parking - Abandoned Vehicle	POLICE	{'latitude': '37.79239919111185', 'longitude': '-122.1993066007311	PENDING		6070673	2115450	CCD4	25X	3500 LOMA VISTA AVE	Oakland	CA	1492
9	967623	2019-12-05T21:29:23.000	SeeClickFix	Illegal Dumping - debris, appliances, etc	ILLDUMP	{'latitude': '37.80189999992631', 'longitude': '-122.2447860008829	PENDING		6057598	2119150	CCD2	15X	2319 PARK BLVD	Oakland	CA	1483
10	967622	2019-12-05T20:48:53.000	SeeClickFix	Illegal Dumping - debris, appliances, etc	ILLDUMP	{'latitude': '37.74411999986224', 'longitude': '-122.1677280002122	PENDING		6079483	2097709	CCD7	32X	9907 WALNUT ST	Oakland	CA	1481
11	967621	2019-12-05T20:45:25.000	SeeClickFix	Illegal Dumping - debris, appliances, etc	ILLDUMP	{'latitude': '37.74381204002426', 'longitude': '-122.1660816080577	PENDING		6075606	2097593	CCD7	32X	9928 WALNUT ST	Oakland	CA	1481
12	967620	2019-12-05T20:24:11.000	SeeClickFix	Parking - Enforcement	PARKING	{'latitude': '37.75003570666936', 'longitude': '-122.1708628277594	PENDING		6078615	2099880	CCD7	34X	1626 92ND AVE	Oakland	CA	1481
13	967619	2019-12-05T19:50:17.000	SeeClickFix	Illegal Dumping - debris, appliances, etc	ILLDUMP	{'latitude': '37.82625443187529', 'longitude': '-122.2819657405531	PENDING		6047027	2128220	CCD3	07X	1200 34TH ST	Oakland	CA	1485
14	967618	2019-12-05T19:39:47.000	SeeClickFix	TE - Traffic Safety (non-emergency)	TRAFFIC_ENGIN	{'latitude': '37.80379732672938', 'longitude': '-122.2462570207995	PENDING		6057192	2119849	CCD2	15X	BROOKLYN AVE	Oakland	CA	1483
15	967617	2019-12-05T19:36:21.000	SeeClickFix	Parking - Abandoned Vehicle	POLICE	{'latitude': '37.75428899871714', 'longitude': '-122.1588799993565	PENDING		6079216	2101417	CCD7	34X	8725 OLIVE ST	Oakland	CA	1493
16	967616	2019-12-05T19:34:36.000	SeeClickFix	TE - Traffic Safety (non-emergency)	TRAFFIC_ENGIN	{'latitude': '37.77091199880885', 'longitude': '-122.2007804993100	PENDING		6070105	2107633	CCD6	27Y	PRINCETON ST & FAIRFAX AVE	Oakland	CA	1188
17	967615	2019-12-05T19:28:12.000	SeeClickFix	Parking - Abandoned Vehicle	POLICE	{'latitude': '37.7414172449819', 'longitude': '-122.1473026116073	PENDING		6085062	2096021	CCD7	32Y	DURANT AVE	Oakland	CA	1109
18	967614	2019-12-05T18:47:12.000	SeeClickFix	Parking - Abandoned Vehicle	POLICE	{'latitude': '37.8481333315140', 'longitude': '-122.2823500013590	PENDING		6047068	2136187	CCD1		6460 ESSEX ST	Oakland	CA	1494
19	967613	2019-12-05T18:44:22.000	SeeClickFix	Streets/Sidewalks - Curb & Gutter Repair	STREETSW	{'latitude': '37.82330999964713', 'longitude': '-122.2009785010499	PENDING		6070394	2126712	CCD4	13Z	5833 SCARBOROUGH DR	Oakland	CA	1488
20	967612	2019-12-05T18:24:12.000	SeeClickFix	Parking - Enforcement	PARKING	{'latitude': '37.83070909608818', 'longitude': '-122.2116351116128	PENDING		6067366	2129162	CCD4	13Z	5500-5758 CABOT DR	Oakland	CA	1488
21	967611	2019-12-05T18:14:10.000	SeeClickFix	Parking - Enforcement	PARKING	{'latitude': '37.77609467068493', 'longitude': '-122.2332375297079	PENDING		6060760	2109694	CCD5	20X	2857 E NINTH ST	Oakland	CA	1188
22	967610	2019-12-05T18:05:10.000	SeeClickFix	Illegal Dumping - debris, appliances, etc	ILLDUMP	{'latitude': '37.78607179941039', 'longitude': '-122.2413486245557	PENDING		6058484	2113570	CCD2	19X	19TH AVE & SOLANO WAY	Oakland	CA	1483
23	967609	2019-12-05T18:03:41.000	SeeClickFix	Parking - Abandoned Vehicle	POLICE	{'latitude': '37.78975335950603', 'longitude': '-122.1980870792631	PENDING		6071006	2114408	CCD4	25X	4000 REDDING ST	Oakland	CA	1492
24	967608	2019-12-05T17:14:34.000	SeeClickFix	Litter - Street Litter Container - Broken	ILLDUMP	{'latitude': '37.81559730620782', 'longitude': '-122.2081457618483	PENDING		6042280	2124420	CCD3	05Y	FRONTAGE RD	Oakland	CA	1484
25	967607	2019-12-05T17:12:52.000	SeeClickFix	Illegal Dumping - debris, appliances, etc	ILLDUMP	{'latitude': '37.76089584680', 'longitude': '-122.2184713410578	PENDING		6065059	2111563	CCD5	24X	1000â€"1014 56TH AVE	Oakland	CA	1188
26	967606	2019-12-05T17:10:52.000	Phone or Ema	TE - Traffic Safety (non-emergency)	TRAFFIC_ENGIN	{'latitude': '37.78574719087895', 'longitude': '-122.2354301150654	WOCREATE		6060192	2113220	CCD2	19X	FOOTHILL BLVD & MUNSON WY	Oakland	CA	1483
27	967605	2019-12-05T17:09:12.000	SeeClickFix	Graffiti on OTHER	OTHER	{'latitude': '37.7955317459747', 'longitude': '-122.2388654816991	PENDING	AT&T,	6059266	2116838	CCD2	17Y	2240 13TH AVE	Oakland	CA	1483
28	967604	2019-12-05T17:09:03.000	SeeClickFix	Other City Services (app/web request)	OTHER	{'latitude': '37.81034768400375', 'longitude': '-122.2761478396432	PENDING		6048597	2122597	CCD3	05X	1855â€"1899 BRUSH ST	Oakland	CA	1489
29	967603	2019-12-05T17:08:42.000	SeeClickFix	Illegal Dumping - debris, appliances, etc	ILLDUMP	{'latitude': '37.78943399988513', 'longitude': '-122.2056030017611	PENDING		6068817	2114404	CCD4	24Y	3500 PENNIMAN AVE	Oakland	CA	1492
30	967602	2019-12-05T17:08:02.000	SeeClickFix	Graffiti on OTHER	OTHER	{'latitude': '37.7910336706622', 'longitude': '-122.2430926041354	PENDING		6057436	2115197	CCD2	19X	1319 FOOTHILL PL	Oakland	CA	1483
31	967601	2019-12-05T17:06:43.000	SeeClickFix	Graffiti on Private Property	GRAFFITI	{'latitude': '37.79106178077', 'longitude': '-122.2471056484080	PENDING		6056854	2115218	CCD2	19X	1205 E 15TH ST	Oakland	CA	1483
32	967600	2019-12-05T17:05:55.000	SeeClickFix	Graffiti on OTHER	OTHER	{'latitude': '37.79418500448916', 'longitude': '-122.2506183359745	PENDING		6055861	2116674	CCD2	19X	735 E 15TH ST	Oakland	CA	1483
33	967599	2019-12-05T17:03:02.000	SeeClickFix	Graffiti on a City Building (Library, Rec Cer	BLDGMAINT	{'latitude': '37.79776128881499', 'longitude': '-122.2633366225413	PENDING		6052211	2117746	CCD2	03X	900 FALLON ST	Oakland	CA	1484
34	967598	2019-12-05T17:02:22.000	SeeClickFix	Graffiti on OTHER	OTHER	{'latitude': '37.79749842555922', 'longitude': '-122.2528165941103	PENDING	USPS,	6052359	2117646	CCD2	03X	43â€"47 TENTH ST	Oakland	CA	1484
35	967597	2019-12-05T17:01:20.000	SeeClickFix	Graffiti on OTHER	OTHER	{'latitude': '37.79754917439179', 'longitude': '-122.2549605223784	PENDING		6051740	2117675	CCD2	03X	55â€"99 NINTH ST	Oakland	CA	1484
36	967596	2019-12-05T16:34:40.000	SeeClickFix	Illegal Dumping - debris, appliances, etc	ILLDUMP	{'latitude': '37.76095300048710', 'longitude': '-122.1790300999999	PENDING		6070325	2103697	CCD6	30X	7500 LOCKWOOD ST	Oakland	CA	1490
37	967595	2019-12-05T16:17:51.000	Phone	Parking - Abandoned Vehicle	POLICE	{'latitude': '37.78630385279085', 'longitude': '-122.2427527422416	OPEN		6058080	2113464	CCD2	19X	1800 E 12TH ST	Oakland	CA	1483
38	967594	2019-12-05T16:17:42.000	Phone	Parking - Abandoned Vehicle	POLICE	{'latitude': '37.78630885279085', 'longitude': '-122.2427527422416	OPEN		6058080	2113464	CCD2	19X	1800 E 12TH ST	Oakland	CA	1483
39	967593	2019-12-05T16:12:55.000	Phone	Tree - Broken/Hanging Limb	TREES	{'latitude': '37.84110161818091', 'longitude': '-122.2391103453341	OPEN		6059503	2133592	CCD1	13X	5805 ROMANY RD	Oakland	CA	1491
40	967592	2019-12-05T16:10:17.000	Phone	Graffiti on Street, Street Light, Traffic Sign	GRAFFITI	{'latitude': '37.83582918887711', 'longitude': '-122.2515932777451	OPEN		6055863	2131540	CCD1	12Y	5211 BROADWAY	Oakland	CA	1491

American Public Service dataset 1(b):



Looking at the dataset on the previous slides, please consider the following questions (5mins).

Where might this dataset come from? Or who might be the owner of this dataset?

What might the dataset's purpose be?

For 2 or 3 of the variables from either this slide (a) or the next (b), write down some ideas of what these variables might mean. Why might these be logged? Who are these variables useful for? Why are they stored the way they are?

Varaiables (a)

Varaiables (b)

For this exercise, you'll be picking 2 or 3 rows from the dataset and jotting down some ideas of how the call might have gone. I've put some post-it notes down below to get you started (15-20mins)

Workshop five (WS5)

Instructions

Today, we'll be rapid prototyping an algorithm. We're going to go through this in the following steps;

- brainstorming potential project ideas
- mapping out your stakeholders
- deciding on your data
- deciding on assessment and documentation standards

This should take approximately a little over an hour. Once you have finished prototyping your algorithm, we'll take a quick comfort break of 5mins, and then you will present a business case for the algorithm. The presentation should last approximately 10mins, and include a detailed business case including how the project will meet ethical standards.

Take 10mins to jot down any work problems you believe it would be beneficial to either automate using algorithmic methods, or would benefit from algorithmic insights. Try and come up with ideas which include a variety of your experiences as a group.

Thinking back to the values and stakeholders' workshop, it's time to think about who the direct and indirect stakeholders might be for this algorithm. Take 10mins to map the indirect and direct stakeholders, including information about in what ways they may be affected and how this could be investigated prior to development.

Now we've thought about who the stakeholders here might be, let's think about how they might best be engaged with. Take 15mins to note down the following things; how might you contact them? How might you find out more about their needs in this project? How might you document this? How might you manage tensions which come up between these stakeholders?

Take 10mins to consider what sort of data this algorithm would need to fulfil this task. This might take different forms, such as thinking about what or who the data would be about, what sort of data is readily available about this already, what these might look like at a variable level. At the end of this task, you should have a list of variables alongside post-it notes detailing how this data would be collected.

Now you have some idea of what sort of data you'll be collecting, it would be good to consider what sort of documentation the project is going to need. Will you do an impact assessment? How will you document your design, data collection, and stakeholder decisions during the project? Who will these be available to? Take 10mins to think these through, and note them on the post-it notes below.

Now you've thought through these issues, it's time to take 10mins to consider how you might present a business case, which includes how this will meet ethical standards, for your project. This doesn't need to be anything new, but a space is provided below incase you wanted to make notes before presenting your idea.

Workshop six (WS6)

Instructions

In today's session, we're going to look at prototyping our framework specific for a DWP context. To do so, we're going to look at deconstructing a few other frameworks first. For the first part of this task, each individually pick one of the frameworks below and take 15-20mins to read or skim through the framework. No worries if you pick a longer one, just get through what you can. Once you've done this, they'll be questions relating to the frameworks on the next few slides, designed to help you break them down. Use these questions to discuss the differences between the frameworks you each read.

PwC's Responsible AI:
https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html

Assessing and addressing algorithmic bias in practice:
https://interactions.acm.org/archive/view/november-december-2018/assessing-and-addressing-algorithmic-bias-in-practice

The Government's Landscape Summary on Bias in Algorithmic Decision Making:
https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819055/Landscape_Summary_-_Bias_in_Algorithmic_Decision-Making.pdf

The UK Government's Ethics, Transparency and Accountability Framework for Automated Decision Making:
https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making/ethics-transparency-and-accountability-framework-for-automated-decision-making

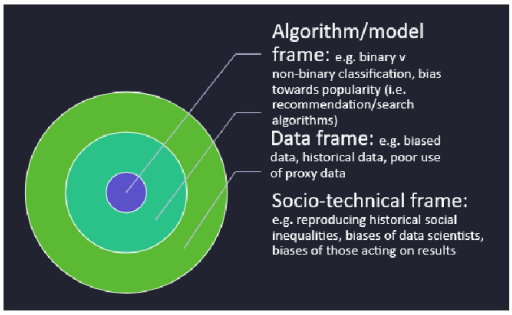
OECD principle's of AI:
https://oecd.ai/en/ai-principles

In the next few slides, we're going to think about deconstructing these frameworks, and think about what each of them offers. There are no right answers here, and all contributions about why these may or may not be useful will be useful for stage 2 - constructing your own framework.

For the following section, we're going to explore what jumps out at you from these frameworks. Consider the following questions: What would you say their core themes are? Who do you think they've been written by? for? What do you think the specific goals of these frameworks are?

Remember back in the introduction, we introduced this model for understanding how algorithmic bias occurs. We have the more technical aspects of a model, such as those in the algorithm and data frame, the social technical aspects such as reproducing historical biases and biases which are introduced by the biases of those designing the algorithm.

On the next slide, we're now going to look at trying to categories some of the elements of the frameworks you've been looking at. We'll also be looking at another category, which is the 'social frame', which includes elements such as policy and assessments which algorithmic technologies sit within, even if they aren't specific to the algorithm itself.



Below are 3 post-it notes, Technical, Sociotechnical, and Social. Which elements you've drawn from the frameworks you've looked at fit in these categories? Where do they fit? Which don't feel like they fit very well?

As you might have realised, these frameworks leave a lot of gaps when thinking about using them for a specific project or organisation. Take a few minutes to think of what is missing from these frameworks for a DWP context, and put them on post-it notes on the space below.

Stage 2

Now we're going to think about what a governance framework for algorithmic projects at DWP might look like. Thinking back to workshop 2 (values and stakeholders), what values and types of action should the framework uphold? And for which stakeholders are these attached to?

Now we've considered the values at play here, we're going to consider how these relate to organisational risks. Take a few minutes to consider what specific risk areas there might be for algorithmic projects at DWP and note them down below. Risk areas can be as specific or as high level as you feel is relevant.

Now we've considered the specific risk areas the framework would need to cover, let's consider what sort of situations it's envisioned this framework would provide support and structure in. Below, take a few minutes to jot down a few situations where an algorithmic bias mitigation framework might be useful. Thinking back to our discussion of values, consider what sort of tensions might exist amongst stakeholders in these situations.

For the final task, look back through your previous answers and consider the following; who would be responsible for ensuring the values of the framework would be upheld? What actions would need to be taken by these responsible parties to ensure these values are upheld? How do these relate to the risk areas you've highlighted?

Workshop seven (WS7)

Instructions

Today, we'll be rapid prototyping an algorithm. We're going to go through this in the following steps;

- brainstorming potential project ideas
- mapping out your stakeholders
- deciding on your data
- deciding on assessment and documentation standards

This should take approximately a little over an hour. On the next 5 slides, there will be questions designed to help you think through each of these stages together as a group. If you have any questions, please don't hesitate to ask.

Take a few minutes to jot down any work problems you believe it would be beneficial to either automate using algorithmic methods, or would benefit from algorithmic insights. Try and come up with ideas which include a variety of your experiences as a group.

Thinking back to the values and stakeholders' workshop, it's time to think about who the direct and indirect stakeholders might be for this algorithm. Take a few minutes to map the indirect and direct stakeholders, including information about in what ways they may be affected and how this could be investigated prior to development.

Now we've thought about who the stakeholders here might be, let's think about how they might best be engaged with. Take a few minutes to note down the following things; how might you contact them? How might you find out more about their needs in this project? How might you document this? How might you manage tensions which come up between these stakeholders?

Take a few minutes to consider what sort of data this algorithm would need to fulfil this task. This might take different forms, such as thinking about what or who the data would be about, what sort of data is readily available about this already, what these might look like at a variable level. At the end of this task, you should have a list of variables alongside post-it notes detailing how this data would be collected.

Now you have some idea of what sort of data you'll be collecting, it would be good to consider what sort of documentation the project is going to need. Will you do an impact assessment? How will you document your design, data collection, and stakeholder decisions during the project? Who will these be available to? Take a few minutes to think these through, and note them on the post-it notes below.

Appendix: C – Workshop outlines from my proposal document to the DWP

Workshop Session 1

Outline: To kick off the workshop series, the first workshop session will include a talk on data-driven discrimination, focusing on the research undertaken as part of projects 1 and 2 of my PhD programme. The talk will also include information about the research being done as part of project 3, how attendees can get involved, and how the findings generated aim to generate a discrimination mitigation framework specifically tailored to the DWP context.

Proposed activities:

- A talk on the findings generated from projects 1 and 2 and how to get involved in project 3
- A question-and-answer session and discussion.

Workshop Session 2

Outline: In session 2, we'll be undertaking a series of thought-provoking exercises around value sensitive design and stakeholder analysis. Using techniques based on previous interdisciplinary research, this workshop aims to help participants reflect on how values are embedded within design practices and recognise how this happens. This will generate knowledge around organisational practices and values, and how these can be linked to discrimination mitigation practices.

Proposed Activities:

- A small talk on value sensitive design, and how values are embedded within all tools and technologies.
- A stakeholder analysis activity which encourages participants to reflect on stakeholder relationships as well as looking further afield to how technology can impact upon indirect stakeholders. This is to encourage broader societal thinking, and ecosystem consideration and management.
- Afterwards, there will be a value source analysis exercise, which will encourage participants to reflect on the projects' values, organisational values, personal values, and the values of direct and indirect stakeholders.

Workshop Session 3

Outline: In this workshop, participants will complete and assess new data-driven impact assessments, giving participants the opportunities to assess prevailing data-driven discrimination themes such as transparency, audits, and standardisation in the context of data-driven technologies. This will generate knowledge as to how organisations and individuals make sense of the administrative elements of data-driven technologies.

Proposed activities:

- This workshop will start off with a 20min talk on developments within organisational impact assessments for data-driven technologies.

- Depending on the number of participants, different groups will be given a different impact assessment to complete for a fictional data-driven system.
- Once each group has reflected on their impact assessment we'll come back to discuss their findings

Workshop Session 4

Outline: In this session, we'll be building on the previous skills by expanding our consideration of data-driven systems. We'll be using story-telling methods to bring the people behind datasets to life, and to consider how technologies play a pivotal role in other people's life courses through the use of ethical vignettes. This will generate knowledge as to how teams make sense of the human side of data and technologies when trying to mitigate the risks of discrimination.

Proposed Activities:

- The first 10mins will include a brief overview of the session.
- Participants will be encouraged to discuss a couple of ethical vignettes which focus on fictional work and welfare data-driven technologies, and asked to reflect on how they would manage the impact of these technologies.
- The groups will then present their reflections back at the end of the session

Workshop Session 5

Outline: This session will bring together the skills and strategies from previous sessions, and give participants the opportunity to design their own data-driven technology and safeguards against discrimination. From a research standpoint this session will provide rich data as to how teams make sense of the mitigation efforts discussed in the previous sessions, and learn valuable insights as to how these methods work in the moment.

Proposed Activities:

- In the first activity participants will be asked to brainstorm as many ideas as possible for a beneficial data driven system.
- Participants will be given an hour to prototype a data driven system. Participants will be asked to either write, draw, or both, a guide which explains what the system does, where it gets its data, what precautions are in place, and how data and discrimination consideration are treated throughout the projects lifecycle.
- A template on a jam board will include all of these categories, as well as an impact assessment for the data-driven system.
- Assuming there are two groups, then the groups will be asked to present their ideas to each other at the end of the session, followed by a discussion.

Workshop Session 6

Outline: The sixth session will wrap up the series with a framework prototyping session. In this session participants will create their own DWP specific framework for mitigating the risks of data-driven discrimination. The outputs from this session, combined with the

knowledge generated in previous sessions, will be used to provide DWP with a framework which brings together all these understandings.

Learning Activities:

- The session will begin with a short 15min talk, covering some of the latest research into data-driven ethical frameworks from a number of sources.
- Participants will reflect on their experiences from the previous sessions, and be asked to use their unique perspectives at DWP to mock-up their own ethical framework for data-driven technologies at DWP

Appendix: D – Ethics approval from the University of Sheffield

The University Of Sheffield.

Downloaded: 23/10/2023
Approved: 27/01/2020

Hadley Beresford
Registration number: 180116133
Sheffield Methods Institute
Programme: Yr2 Data Analytics and Society

Dear Hadley

PROJECT TITLE: Mapping Algorithmic Bias
APPLICATION: Reference Number 032329

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 27/01/2020 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 032329 (form submission date: 23/01/2020); (expected project end date: 01/05/2020).
- Participant information sheet 1073985 version 3 (23/01/2020).
- Participant information sheet 1073984 version 3 (23/01/2020).
- Participant consent form 1073986 version 2 (23/01/2020).

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Your responsibilities in delivering this research project are set out at the end of this letter.

Yours sincerely

Todd Hartman
Ethics Administrator
Sheffield Methods Institute

Please note the following responsibilities of the researcher in delivering the research project:

- The project must abide by the University's Research Ethics Policy: https://www.sheffield.ac.uk/research-services/ethics-integrity/policy
- The project must abide by the University's Good Research & Innovation Practices Policy: https://www.sheffield.ac.uk/polopoly_fs/1.671066!/file/GRIPPolicy.pdf
- The researcher must inform their supervisor (in the case of a student) or Ethics Administrator (in the case of a member of staff) of any significant changes to the project or the approved documentation.
- The researcher must comply with the requirements of the law and relevant guidelines relating to security and confidentiality of personal data.
- The researcher is responsible for effectively managing the data collected both during and after the end of the project in line with best practice, and any relevant legislative, regulatory or contractual requirements.

The University Of Sheffield.

Downloaded: 23/10/2023
Approved: 08/03/2021

Hadley Beresford
Registration number: 180116133
Sheffield Methods Institute
Programme: SMI-Data Analytics & Society CDT

Dear Hadley

PROJECT TITLE: Exploring attempts to mitigate bias in public services
APPLICATION: Reference Number 038401

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 08/03/2021 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 038401 (form submission date: 17/02/2021); (expected project end date: 31/07/2021).
- Participant information sheet 1087698 version 1 (17/02/2021).
- Participant consent form 1087699 version 1 (17/02/2021).

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Your responsibilities in delivering this research project are set out at the end of this letter.

Yours sincerely

Joanna Eve
Ethics Administrator
Faculty of Social Sciences

Please note the following responsibilities of the researcher in delivering the research project:

- The project must abide by the University's Research Ethics Policy: https://www.sheffield.ac.uk/research-services/ethics-integrity/policy
- The project must abide by the University's Good Research & Innovation Practices Policy: https://www.sheffield.ac.uk/polopoly_fs/1.671066!/file/GRIPPolicy.pdf
- The researcher must inform their supervisor (in the case of a student) or Ethics Administrator (in the case of a member of staff) of any significant changes to the project or the approved documentation.
- The researcher must comply with the requirements of the law and relevant guidelines relating to security and confidentiality of personal data.
- The researcher is responsible for effectively managing the data collected both during and after the end of the project in line with best practice, and any relevant legislative, regulatory or contractual requirements.

The University Of Sheffield.

Downloaded: 23/10/2023
Approved: 03/03/2022

Hadley Beresford
Registration number: 180116133
Sheffield Methods Institute
Programme: SMI-Data Analytics & Society CDT

Dear Hadley

PROJECT TITLE: Algorithmic bias in the public sector
APPLICATION: Reference Number 045062

On behalf of the University ethics reviewers who reviewed your project, I am pleased to inform you that on 03/03/2022 the above-named project was **approved** on ethics grounds, on the basis that you will adhere to the following documentation that you submitted for ethics review:

- University research ethics application form 045062 (form submission date: 24/02/2022); (expected project end date: 31/05/2022).
- Participant information sheet 1101670 version 3 (24/02/2022).
- Participant information sheet 1101669 version 3 (24/02/2022).
- Participant consent form 1101673 version 2 (24/02/2022).
- Participant consent form 1101672 version 2 (24/02/2022).

If during the course of the project you need to deviate significantly from the above-approved documentation please inform me since written approval will be required.

Your responsibilities in delivering this research project are set out at the end of this letter.

Yours sincerely

Karen Bralsford
Ethics Administrator
Sheffield Methods Institute

Please note the following responsibilities of the researcher in delivering the research project:

- The project must abide by the University's Research Ethics Policy: https://www.sheffield.ac.uk/research-services/ethics-integrity/policy
- The project must abide by the University's Good Research & Innovation Practices Policy: https://www.sheffield.ac.uk/polopoly_fs/1.671066!/file/GRIPPolicy.pdf
- The researcher must inform their supervisor (in the case of a student) or Ethics Administrator (in the case of a member of staff) of any significant changes to the project or the approved documentation.
- The researcher must comply with the requirements of the law and relevant guidelines relating to security and confidentiality of personal data.
- The researcher is responsible for effectively managing the data collected both during and after the end of the project in line with best practice, and any relevant legislative, regulatory or contractual requirements.