A critical exploration of ethical discourses around the adoption of an algorithmic tool by a criminal justice case study in Europe: A Foucauldian perspective

Ali Gordjahanbeiglou

Doctor of Philosophy

ASTON UNIVERSITY

September 2024

ASTON UNIVERSITY

A critical exploration of ethical discourses around the adoption of an algorithmic tool by a criminal justice case study in Europe: A Foucauldian perspective

Ali Gordjahanbeiglou

Doctor of Philosophy

2024

THESIS SUMMARY

As the use of algorithmic technologies for key organisational and work processes grows, the AI/algorithm ethics literature also continues to raise important concerns around the potential moral risks associated with their use, such as the unintended biases that may stem from algorithmic decision-making. More recently, research has begun to highlight the roles of human self-reflexivity and resistance, calling for further research into this burgeoning stream of AI/algorithm ethics. This thesis, therefore, builds upon these underexplored ethical nuances in algorithmic work practice utilising a Foucauldian lens – in particular, drawing Foucault's theories of *discourse*, *governmentality,* and *resistance/ethics* – to explore the ethical discourses and actions that emerge when a large/complex criminal justice organisation (based in a European country) adopted algorithmic tools to aid in their key decision-making activities.

Data was collected through 38 semi-structured qualitative interviews with different organisational actors. A range of organisational documents were used as an addition to the interview data. Using Foucauldian Discourse Analysis theory, I have found that the adoption of algorithmic technologies in this particular service was steered and supported by the scientific power/knowledge of data scientists. I also found that whilst transparent (and ethical) work practice, for data scientists and senior leaders, is achieved via utilisation of algorithms and data-driven tools, there is a nascent discursive shift amongst many frontline practitioners. This discursive shift highlights practitioners' agency, self-reflexivity and awareness around shortcomings and potential ethical risks of algorithms. I argue that the practitioners' awareness and – in some cases – subtle resistances against algorithm are examples on how ethical practice is crystalised in algorithmic work environments. By applying a Foucauldian lens, this thesis contributes to organisational ethics and AI/algorithm ethics literatures, highlighting ethical nuances in relations to marginalisation of employee voices (discourses) through algorithmic work governmentality. Moreover, this research gains further understanding on how those marginalised discourses shift towards subtle active resistance and expansion of the space for ethical practice.

Keywords: AI, Algorithmic work practice, Ethical discourses, Foucauldian perspective, Governmentality, Criminal justice organisation

DEDICATION

To all women in Iran who fight for equal rights

"Woman, Life, Freedom"

Signature slogan, 2022 Mahsa civil movement

ACKNOWLEDGEMENTS

It would have been impossible to complete this PhD thesis without the support of so many people, for which I will always be grateful. First, I would like to thank my supervisory team, Dr Jonathan Crawshaw, Dr Judy Scully and Professor Nick Theodorakopoulos. Jonathan and Nick always encouraged me to produce my best. Their comments have always been beneficial and constructive. I would also like to thank Judy for her unlimited support. It was her warmth, wisdom, fortitude that inspired me. I can never thank you enough for the constant encouragement throughout all these years.

I would also like to thank Aston University Business School and the Department of Work and Organisation. Especially, I would like to thank Dr David Cantliff and Dr Kanimozhi Narayanan for providing me with the opportunity to deliver some seminars in their modules. I also thank the criminal justice service and all the participants who agreed to take part in this study. Your contributions to this research are greatly appreciated and it was a pleasure to meet you all. I would also like to thank the BSS director for PhD programmes at Aston University, Dr Vahid Sadeghi, for his invaluable advice on publications and careers in academia. I would also like to express my gratitude to my viva voce examiners, Professor Pawan Budhwar and Professor Andrew Kakabadse for their time and invaluable feedback.

I would also like to thank my PhD friends and colleagues at 11th floor, Aston University Main Building. I had a nice time being there and it was such an honour to work alongside you. I am going to miss all our ‘little academic debates’ and ‘free lunches’ together.

I would also like to express how grateful I am to my partner and the love of my life, Rebecca. I met you during this PhD journey and your patience, support and encouragement made it possible for me to complete this thesis. You were always there for me with your abundant love. I wish our paths would have crossed sooner.

Above all, I would like to express my sincere gratitude to my family back home, in particular, to my parents. Although we live quite far away from each other, you have always been there for me and supported me in every step of the journey in so many ways. I appreciate all the sacrifices you made for me, and everything you have done that made it possible for me to work on the PhD. I am lucky to have you and can never thank you enough. Finally, my thanks to anyone who takes the time to read my work. I hope you find it useful and informative.

PRESENTATION AT CONFERENCES

Conference	Paper
40th EGOS Colloquium, University of Milano-Bicocca (July 4-6, 2024), Milan, Italy	Ethical discourses around the implementation and utilisation of algorithms in a criminal justice system: A Foucauldian perspective
1st Interdisciplinary PhD conference, College of Business and Social Sciences, (June 2022), Aston University, Birmingham	Adoption of Artificial Intelligence (AI) in Organisational work practices: An Exploration!

LIST OF CONTENTS

LIST OF ABBREVIATIONS

AI: Artificial Intelligence

ARA: Algorithmic Risk Assessment

CAQDAS: Computer-assisted Qualitative Data Analysis Software

CEO: Chief Executive Officer

COVID-19: Coronavirus Disease 2019

DOI: Diffusion of Innovation [model]

ESRC: Economic and Social Research Council

FDA: Foucauldian Discourse Analysis

GDPR: General Data Protection Regulation

HRM: Human Resource Management

ICT: Information and Communication Technology

IT: Information technology

TAM: Technology Acceptance Model

TOE: Technology, Organisation, Environment [model]

UK: United Kingdom

LIST OF TABLES AND FIGURES

CHAPTER 1: Introduction

1.1 Background and Context of Research

Artificial Intelligence (AI) is an umbrella term that includes many intelligent technologies, including algorithms. The concept of AI has been a controversial topic in public discourse for many years, often portrayed as sentient entities in sci-fi motion pictures with the aim to dominate the world and make the human race either enslaved or annihilated. This depiction of AI is rather unrealistic and comical, as artificial intelligence technologies have been with us for several years, and the majority of us interact with such technologies on a daily basis. From statistical algorithms utilised for workforce management to sophisticated AI technologies such as Watson made by IBM, which is based on natural language processing (Jiang et al., 2017), or the AI designed for autonomous vehicles (Manfreda et al., 2019), AI and algorithmic technologies no longer exist in the realm of science fiction or futurologists, but they are staple parts of the structure for many organisations (Dwivedi et al., 2021). Organisations are always in need of strategies, tactics, and solutions to boost productivity, service quality, and saving costs. Since early 2020 and the start of the COVID-19 crisis, safety measures have significantly affected work environments, forcing many jobs into remote work fashions (Leonardi, 2021). This crisis has paved the way for further utilisation of AI and other algorithmic technologies in order to overcome the work challenges of the pandemic (see: Chowdhury et al., 2023; Suseno et al., 2021; Makarius et al., 2020). So far, algorithmic technologies have come with several promises, such as better data processing (Holford, 2019) and enhanced decision-making (Lee and Shin, 2020) for organisational work processes. This technological transformation due to the rise of AI has led to substantial enthusiasm within academia to better understand and evaluate the benefits, impacts, and consequences of adopting these intelligent agents in organisations.

At the moment, there is no universal definition for an algorithm in the existing literature (Tsamados et al., 2022). For instance, Hill (2016) defines an algorithm as *"a finite, abstract, effective compound control structure, imperatively given, accomplishing a given purpose under given provisions"* (p. 44). Whereas Lindebaum et al. (2020) see algorithms as computer-based rules or calculations for automated decision-making and/or problem-solving. Inspired by the optimisation capabilities of algorithms, Zhou et al. (2022) depict algorithms as powerful computational systems able to make decisions based on rules and mathematical models applied to the evaluation of available data. Kraemer et al. (2010) describe algorithms as computerised entities implemented to solve a wide range of problems. As explained, there are several insights around the concept of AI, but Kraemer et al.'s (2010) insight seems more applicable to the context of this research.

The premise of AI has often appeared in the literature surrounding ‘immersive technologies’ (Dwivedi et al., 2022; Wei and Yuan, 2023). Immersive technologies are considered those that can bridge the gap between the physical and virtual worlds (Suh and Prophet, 2018), creating a sense of immersion for the user within the virtual environment (Wei and Yuan, 2023). It means that immersive technologies enable the user to interact and engage in the virtual space and fulfil corresponding conducts and functions (Baxter and Hainey, 2024). The idea of immersive technologies is leveraged by the existence of tools such as augmented reality (AR) and virtual reality (VR) (Tom Dieck and Han, 2022). Although AI and algorithmic tools can widely be utilised to enhance the immersive technologies’ functionality and user experience (Xi et al., 2024), they cannot be categorised as solely immersive technologies. This is because AI and/or algorithms lack sensory-rich interactive features (Nilashi and Abumalloh, 2024) that are essential components of immersive technologies. Given the significant differences between algorithmic technologies and immersive ones, it is sensible to incorporate the term ‘algorithm’ when referring to the technology. Furthermore, given the advice provided by the criminal justice case study of this research regarding the predictive data-driven nature of the utilised algorithm tool, I consider this term throughout this thesis.

To date, there has been extensive research on how algorithmic technologies can effectively be designed and introduced within human organisations. As such, there is a robust literature within the realm of computer/data science that has looked at improving machine intelligence (Tweedale, 2013) or how to elevate algorithms’ capabilities as team members within human work assemblages (Seeber et al., 2020). There has even been a significant body of research in organisational behaviour literature that examines the perceptions, work experiences, and behaviours of humans in relation to their interactions with AI and algorithmic technologies (e.g., Pachidi et al., 2021; Leicht-Deobald et al., 2019; Bucher et al., 2021; Sherwani et al., 2020; Bader and Kaiser, 2019; Etter and Albu, 2021).

However, as the literature around the merits of algorithmic technologies in work organisations continues to grow, there is another emerging strand of literature that highlights a darker side of these technologies. In that respect, this strand of literature underscores that algorithms are neither innocuous nor ethically neutral entities (Floridi and Taddeo, 2016). As such, it has been argued that there may be significant ethical risks associated with the utilisation of algorithms for organisational work practices (Tsamados et al., 2022). The premise of ‘ethical AI’ has been developed as a response to ensure that these technologies are aligned with moral principles. In other words, ‘ethical AI’ underscores the moral obligations and duties of an AI and its creators to ensure ‘good’ or ‘right’ decisions and actions are made (Siau and Wang, 2020). The existing literature has categorised the

algorithm ethics in relation to the issues that emerge during the design and development stage (e.g., human biases that affect data as well as data privacy and transparency issues) (Mittelstadt et al., 2016) and the ethical concerns caused by algorithms' outputs, including unfair outcomes and loss of human agency (Peeters, 2020). This body of research – particularly, within the computer/data science discipline – has highlighted concepts of accountability, explainability, and transparency as a panacea to overcome the ethical issues of algorithm technologies (Hoffmann et al., 2018; Kim et al., 2020). Such roadmaps can lead to the development of algorithms that are less 'opaque' (Heßler et al., 2022) and mitigate the inscrutability issue of these tools (Waardenburg et al., 2022).

Scholars, however, have criticised the existing ethical guidelines and frameworks (developed by computer science researchers) for AI/algorithms, highlighting that such guidelines may not capture all the nuances of ethics in relation to human-algorithm work interactions. For instance, Ananny and Crawford (2018) argue that algorithm transparency is not solely adequate for the creation of ethical algorithms. They explain that there are several limitations with the idea of transparent algorithms, including the negligence of 'power asymmetries' and 'agency' of individuals. Similarly, Kim and Moon's (2021) research illuminates that although algorithmic transparency research has helped to identify the downsides of cryptic algorithmic outputs, it fails to provide insights into how human end-users perceive the ethical side of algorithmic technologies. As such, according to Kim and Moon (2021), it is beneficial to understand human end-users' perspectives since not only can it provide better explanations for them on how an algorithm operates, but it also helps to deeply understand in-built biases in an algorithm from their view.

The prior research on AI/algorithm ethics, to date, has tended to focus on the ethical aspects of AI/algorithm technology itself, with the aim to ensure that algorithms are scientifically created and designed in an ethical manner (e.g., Tsamados et al., 2022; Mittelstadt et al., 2016; Siau and Wang, 2020; Ouchchy et al., 2020; Etzioni and Etzioni, 2017). Such academic discussions around the ethics of AI technology are essential to combat bias and ensure data privacy (Chowdhury et al., 2023). Yet, making the technology more ethical is only the first step. Whether algorithmic technologies are sought for automation or augmentation of work processes, there are individuals whose working lives are being affected by these transformations (Kellogg et al., 2020). Theorists have argued that there is an emerging dynamic of 'power' (Neyland, 2015) in relation to intelligent algorithms and command-and-control systems of power that affects our working lives (Amicelle, 2022; Anteby and Chan, 2018; Cooper, 2020). There is power invested in algorithms to provide a logic and a 'truth' that drives organisational practices. And this enacted power of algorithms tends to mechanise the human end-user, directing and

controlling their cognition and actions (Peeters, 2020). In a nutshell, algorithms are deemed as agential and powerful technologies that are firstly, difficult to govern (Floridi, 2018), and secondly, 'act on us' (Neyland, 2016). With regards to the concept of ethical AI/algorithms, Neyland (2016) underlines that having an ethical algorithm or holding an algorithm accountable is challenging since algorithms are situated within power asymmetries and associations (Neyland and Möllers, 2016).

However, this specific focus in the literature on the notion of the power of algorithmic technologies seems to be predominantly directed to 'algorithmic surveillance' (Amicelle, 2022) conducted by either state organisations or those with high monetisation goals of 'surveillance capitalism' (Zuboff, 2019). Little is known about how these power dynamics work within other organisational contexts where surveillance is not on the primary agenda. Furthermore, despite this abundance of research on algorithmic surveillance and debates of ethics, the roles and experiences of organisational actors are still ambiguous within these convoluted power dynamics (Introna, 2016). Therefore, It is indeed to understand the behaviours of those organisational actors who are subjected to the agential power of algorithms (De Laat, 2019). This is because the attitudes, perspectives, behaviours, and experiences of organisational actors in response to a technological transformation can fundamentally determine an organisation's readiness for change (Jöhnk et al., 2020). Furthermore, the extent to which organisational actors perceive algorithms usefulness and ease of use are pivotal factors that outline acceptance of, or aversion towards, these tools (Heßler et al., 2022). Thus, the question that is raised here is how organisational actors respond to algorithmic tools. "Do they comply, negotiate, or resist" (De Laat, 2019, p. 326)?

In recent years, the body of research on AI/algorithm ethics has seen a shift from ethics in computer and data science knowledge to critical organisation studies. This shift within scholarship detaches itself from notions such as algorithmic 'in-built' biases and solutions such as algorithmic accountability and transparency. This scholarly shift revolves around human agency, aiming to explain the ethical nuances in relation to human-algorithm interactions (e.g., Introna, 2016; De Vaujany et al., 2021; Newlands, 2021; Weiskopf and Hansen, 2023; Jarrahi et al., 2022). This burgeoning field calls for further research on human intelligence augmentation via algorithms in organisational settings with limited research such, as criminal justice systems (Jarrahi et al., 2022), and to explore people's 'resistances' against these intelligent systems (Newlands, 2021). Morality and ethical practice are integral notions in the ethos of judicial and legal systems (Hazard, 1990). The criminal justice systems are founded on the basis of accountability, impartiality, and transparency (McKay, 2020). And the decisions made in any legal system context have significantly high stakes, determining an individual's liberty or incarceration (Oswald et al., 2018). However, the rise of

AI and algorithms in judicial settings has caused a malaise amongst the criminal justice scholarship (see: Hartmann and Wenzelburger, 2021; Završnik, 2021; Skeem and Lowenkamp, 2020; Schwerzmann, 2021; Simmler et al., 2023). For instance, the issue of the 'black box' nature or inscrutability of algorithms is outlined in the relevant literature, which makes it hard for practitioners to understand algorithmic predictions (McKay, 2020). Moreover, the ethical issues such as biased risk assessment and disproportionate sentencing are also highlighted due to the deployment of algorithmic predictors (Chouldechova, 2017). Thereby, understanding actors' perspectives remains a critical concern in algorithm ethics as undesirable algorithmic outcomes may pose threats to human dignity (Jarrahi et al., 2021) and self-determination (Mittelstadt et al., 2016).

Within this growing field of research, the premise of 'governmentality politicisation' is put forward by Weiskopf and Hansen (2023), which sees algorithms as instruments of power/knowledge constituted via scientific – experts – discourses (Shaw and Scully, 2023). As such, the prevalence of algorithmic work practice is envisaged as a potential threat to human subjectivity and reflexivity (Leonardi and Treem, 2020). In that regard, Zuboff (2019) reveals a new authoritarian governing regime catalysed by the existence of algorithms, targeting human dignity and autonomy. Although, on the surface, it seems that the space for ethical practice is shrinking because algorithms might endanger our autonomy, Weiskopf and Hansen (2023) argue that the space for ethical conduct can be 'opened up.' As such, they underscore that ethical practice is proliferated through new forms of "problematisation, contestation, and resistance" (p. 489) sought and acted upon by different actors located in algorithmic work regimes. In that sense, the actors are capable of 'ethical work,' as they reflect on their work interaction with algorithms and strive to become what or who they wish to be (Crane et al., 2008).

Two significant shortcomings have been identified within the existing literature that this research aims to address. First, the majority of publications that have pioneered the premise of ethical AI and/or algorithms are steered by the computer/data science discipline, offering generic or rather prescriptive solutions for overcoming ethical concerns (e.g., Etzioni and Etzioni, 2017; Tsamados et al., 2022). Indeed, such research is greatly appreciated as it helps to academically build ethical algorithms as well as develop robust ethical toolkits (Floridi et al., 2018). Yet, this stream of research seems to offer only a limited contextualised and empirical basis from organisational environments: As highlighted by Jarrahi et al. (2022), more in-depth research is needed in social environments such as legal or criminal justice systems to understand who benefits and who suffers from algorithmic augmentations. In addition, this strand of research has tended to rely on particular hegemonic discourses of technology experts (mostly data scientists published in academic outlets such as *Big Data &*

Society or similar ones) and lacks empirical inclusivity, particularly from the human end-user dimension. An in-depth understanding of human end-users' perspectives on the ethics of algorithms seems essential as they shape or challenge their interactions with algorithm agents (Burton et al. 2020). Furthermore, a better understanding of human actors' perspectives is the key to exploring resistance in human-algorithm interactions (Kellogg et al., 2020).

Second, there is a growing body of research that sheds light on underexplored notions such as the agential power of algorithms (Neyland and Möllers, 2016) and actors' self-reflexivity and resistance (Weiskopf and Hansen, 2023) within algorithmic work practices. This newer strand of literature on ethics revolves, specifically, around human-algorithm interactions and highlights that algorithmic work regimes do not necessarily impede individuals from ethical conduct. This domain of research also unearths new modes of questioning or problematisation by human end-users who exercise their awareness and reflexivity (Introna, 2016) and take action to disclose the potential threats of algorithmic work regimes. However, this burgeoning field in algorithmic ethics emphasises the need for further empirical evidence: further research that can scrutinise how algorithmic work settings circumvent human reflexivity and how day-to-day activism, and resistance are formed by individuals (Weiskopf and Hansen, 2023). Thereby, this research answers their call for further empirical research on this expanding body of research on AI/algorithm ethics, utilising a Foucauldian lens.

Similar to the shift in AI/algorithm ethics literature, there has been more attention to the perspectives of human professionals in criminal justice organisations in relation to human-algorithm interactions. In this strand of literature, scholars have looked into how algorithms' predictions shape the practices (Hartmann and Wenzelburger, 2021). In light of algorithmic risk predictions, Hartmann and Wenzelburger (2021) demonstrate that algorithmic agential power (Neyland and Möllers, 2016) might affect human practitioners' cognition, tweaking or reconfiguring their agency. In a similar vein, Završnik (2021) argues that the rise of a new algorithmic digital elite in criminal justice systems has become a hindrance to organisational actors' autonomy and agency. The deployment and reliance on algorithmic judicial practices demonstrate the lack of trust the system has in its stakeholders, including judges or parole officers (Završnik, 2021).

The focal attention on the roles and perspectives of organisational stakeholders in criminal justice organisations raises interesting questions that could further expand algorithm ethics literature. An anchor underexplored area relates to the earlier discussion around the agency and activism of relevant organisational actors. As many judicial systems heavily rely

on algorithmic predictions, how is human professionalism directed, tweaked, and/or overshadowed by such technologies? And how do actors exercise their own subjectivity and constitute their centrality within algorithmic practice? It is also essential to explore how civil servants voice concerns, questions, or even resistance regarding the utilisation of technology. These are underexplored – yet crucial – areas that directly affect an organisation's ethos of ethics. As Schwerzmann (2021) argues, the meaning of justice can shift amidst algorithmic decision-making: Empowering/augmenting human assessment with algorithms could jeopardise the stability and integrity of a criminal justice organisation. Thus, it is essential to explore the potential decay of practitioners' autonomy and agency within algorithmic criminal justice systems (Hartmann and Wenzelburger, 2021).

Weiskopf and Hansen (2023) invoke the rather abstract concept of 'governmentality' developed by Foucault to explore ethics in algorithmic work practices. Foucault was predominantly interested in the administration and organisation of people's lives (Townley, 1993) across a range of power/knowledge institutions, including prisons, judicial settings, military barracks, psychiatric clinics, and schooling (Barratt, 2003). To date, his most renowned work on disciplinary power, surveillance, and regulations of individuals has been particularly inspirational amongst critical management scholarship to understand various organisational phenomena (e.g., Hardy and Thomas, 2014; Knights, 2002; Leclercq-Vandelannoitte, 2011; Seeck and Kantola, 2009; Matthewman, 2013). Foucault's seminal works illustrate a depth of analysis into how our lives are situated in different power/knowledge dynamics and how we, as 'subjects,' can become 'docile' or 'aware' of the disciplinary mechanisms of power (McKinlay and Starkey, 1998). Even though Foucault's philosophies were published in an era when AI technology and automation were only hypothetical, they have been incredibly insightful in the context of the 4.0 industrial revolution, the rise of neoliberal capitalism, and the gig economy (De Vaujany et al., 2021; Raffnsøe et al., 2019). This indicates the relevance of his works to the research on AI and algorithm work practices.

Foucauldian theories entail several dimensions, ranging from his early archaeological work, *The Order of Things* (Foucault, 2018), to later genealogical works such as *Discipline and Punish* (Foucault, 1977) and *The History of Sexuality, Vol 1-3* (Foucault, 2020b; Foucault, 2019; Foucault, 2020c). Due to the depth of Foucault's analyses and his profound, constant reflection on (co)production of subjects, it is hard to completely separate one premise from another (Raffnsøe et al., 2019). Yet, for the purpose of this study, the researcher adopts two notions that are most informative to the context. First is the Foucault's particular view on 'discourses', which he conceptualised as avenues of contemplation and communication that form our social interactions in talking/writing and subsequently pave the

way for the presentation/exercise of power (Fairclough, 1993). Discourse, for Foucault, can be an emancipatory phase that 'unfreezes' people thinking to embrace change (Mingers and Willcocks, 2004). Of course, discourse, as Foucault argues, includes – but is not limited to – intangible meanings within enunciation, statement, and writings. Discourse can entail rules, interplays, tactics, and strategies that shape the social milieux (Alvesson and Karreman, 2000) and open spaces for resistance (Foucault, 2020a).

Second, the concept of governmentality, which is a follow-up to Foucault's problematisation on power/knowledge. He developed the notion of governmentality in response to neoliberal governance (Larner, 2000) and brings to light particular rationalities, technologies, and discourses of neoliberal governments that aim to control and objectify people and make 'ideal citizens' (Moisander et al., 2018). The art of governmentality, for Foucault, is not a direct authoritarian control that impedes citizens' autonomy (Ahonen et al., 2014). But it is rather a particular knowledge that neoliberalism instrumentalises to make the population self-governed and self-regulated (Raffnsøe et al., 2019). In other words, governmentality is about governing well by governing less (Mennicken and Miller, 2014).

This research, therefore, posits that there are nuances in algorithmic work practice in the sense that subjects and their expertise can become governable through these technologies (Introna, 2016). Algorithms are agents that aim to tackle shortcomings in human intelligence (Dwivedi et al., 2021), seek to harmonise, normalise, and regulate working lives of other actors in organisations. This aligns with neoliberal governance tendencies that push subordinate power institutions (such as education or crime control settings) towards exercising more control, regulating, and monitoring subjects (Moisander et al., 2018). This intensification of control and normalisation is not solely directed towards governance of people outside of an organisation, as shown in great detail by scholars such as Bakir (2015), Du Plessis (2020), or Zuboff (2019). Rather, algorithms can be deemed novel technologies for exercising governmentality that simultaneously regulate, control, and direct the working lives of organisational actors (Roberts, 2019), whilst enhancing human performativity (Jarrahi et al., 2022).

In the realm of ethics in governmentality through algorithms, Weiskopf and Hansen (2023) discuss that ethical practice is still pervasive despite the dominating power of algorithms. Algorithmic agents seem to govern and direct the conduct of actors. They depict an authentic 'truth' for the users and, to some extent, objectify them (De Laat, 2019). Whilst it seems algorithms lead to marginalisation of human agency and reflexivity (Introna, 2016), ethical conduct stems from people's awareness, contestation, and resistance to algorithmic work governmentality (Leonardi and Treem, 2020). Put differently, algorithms are able to

circumvent or nudge human agency. Research indicates that humans can bring novel forms of questioning and problematisation into their interactions with algorithms (Cameron and Rahman, 2022; De Vaujany et al., 2021). A newer line of thought into algorithmic ethics, proposed by Weiskopf and Hansen (2023), considers how human actors are able to evade or escape algorithmic control using different individual or collective strategies.

Foucault's underexplored theories on ethics (Foucault, 2020c) seem promising and align well with the mentioned line of enquiry into algorithmic ethics. The ethical dimension in Foucauldian philosophies highlights the subject's activism and reflexivity against any modalities of dominance and control (Crane et al., 2008). Ethical subjects, according to Foucault (2020c), are constituted when they think, act upon, and transform themselves to attain the virtue of wisdom, happiness, or freedom. This notion in Foucault's theories requires more reflection and contextualisation in organisational studies (Raffnsøe et al., 2019), specifically when it comes to technologies such as algorithms that might not be ethically neutral (Tsamados et al., 2022).

Ultimately, drawing on Foucauldian theories can aid in theoretically and empirically unearthing those ethical nuances, particularly those relevant to human actors' perspectives and conduct, rather than focusing on the technology itself.

1.2 Research Questions

In order to address the mentioned gaps in the existing knowledge, this study proposes two research questions:

RQ1: What are the dominant ethical discourses around the deployment of algorithms from the perspective of key organisational actors?

RQ2: How do organisational actors influence the ethical discourses of algorithms, and to what extent do they change or challenge their work experience with these tools?

1.3 Research Objectives

The study's objectives are:

RO1: To review the literature pertaining to AI/algorithm ethics and develop a theoretical lens based on Foucault's works.

RO2: To explore the adoption and implementation processes for the used algorithmic tools in the criminal justice case study of this research.

RO3: To understand the dominant discourse in relation to the ethics of algorithmic work practice based on the perspectives of actors at the criminal justice case study.

RO4: To understand human agency and subjectivity in AI/algorithm ethics and further expand that growing body of literature.

1.4 Contributions to theory

This qualitative research makes important contributions to the existing literature on organisational ethics and critical management studies as well as the AI/algorithm ethics literature. First, this research highlights that the art of governmentality in novel neoliberal organisations (Raffnsøe et al., 2019) has become a dominant discourse via the instrumentalisation of algorithmic technologies. In other words, algorithms might become agents that can influence human decision-making (Floridi et al., 2018), which can lead to the marginalisation of human agency and reflexivity as a result of this novel art of governmentality (Introna, 2016). Previous research has considered that the implementation of algorithms may have significant impacts on the behaviours of organisational actors. Whilst acceptance of algorithms by some organisational actors could highlight their conformity and compliance towards governmentality (Pachidi et al., 2021), other actors may contest or resist algorithms (Newlands, 2021). Within the exercise of governmentality, however, 'ethics' remains an underexplored notion, specifically in relation to governmentality through algorithms (Weiskopf and Hansen, 2023). Furthermore, given the importance of ethical, transparent practice in criminal justice organisations (Hazard, 1990) and ethical issues that could affect them due to algorithms (Simmler et al., 2023), the extent to which discourse of algorithm ethics shifts in such organisations is open for further scrutiny.

Previous research has also considered the proliferation of organisational governmentality via the utilisation of algorithms, highlighting their surveillant nature with potential ethical impacts (e.g., Newlands, 2021; Zuboff, 2019; Introna, 2016; Cooper, 2020; Barry, 2019). The strand of research pictures algorithms as novel commodities with significant impacts on human autonomy, agency, and subjectivity (Introna, 2016), as humans become objects or clusters of data to be processed by the machine (Leonardi and Treem, 2020). This study provides empirical and contextualised insights for this growing body of research, unfolding a particular power/knowledge discourse that has catalysed the utilisation of algorithms. This study also illuminates further on algorithmic governmentality (Weiskopf and Hansen, 2023) and brings to light expertise scientific discourses that advocate algorithm and marginalise other actors' voices, including the human end-users. This study, thus, joins the conversations by Larner (2000) and Moisander et al. (2018) by unearthing the *political*

and *disciplinary* elements of governmentality in the empirical context of this research and also pinpoints algorithms as novel instruments in the art of governmentality. By doing so, this research outlines that algorithms are mechanisms that may have fundamental ethical impacts on organisational actors such as exclusions and marginalisation of voices from strategic decision-making (Ford and Harding, 2003) or algorithm design processes (Chatterjee et al., 2021). Therefore, not only does this study provide empirical evidence and contribute to the emerging research on the ethics of 'governmentality through algorithms' (Introna, 2016; Weiskopf and Hansen, 2023), but it also unearths that this form of governmentality is steered by the power/knowledge of data science actors. Using Foucault's concepts, this study contributes to the literature in relation to the ethics of organisations (Chye Koh and Boo, 2004) by highlighting that the exercise of governmentality through algorithmic work practice may have ethical ramifications. In that regard, this study has identified an intensification of scientific power/knowledge (Hardy and Thomas, 2014) that is steered and strategised by only experts (i.e., data scientists as the sole drivers) and has marginalised – and ignored – other actors' inputs, including practitioner end-users of algorithms.

Second, this research contributes to the literature on the ethics of AI/algorithms. According to current studies, several issues might arise due to the adoption and implementation of algorithms: From salient biased and discriminatory algorithm predictions (Mittelstadt et al., 2016; Floridi, 2018) to humans' over-reliance on machine-predicted outputs (Leicht-Deobald et al., 2019) and decay of our self-determination (Tsamados et al., 2022). As such, scholarship warns that potential ambivalences in algorithmic predictions may jeopardise the perceived trustworthiness of an algorithm in the eyes of users (Roßmann et al., 2018) and result in algorithm aversion amongst users (Dietvorst et al., 2015). Whilst the illumination of these ethical issues is sufficiently broad, helping the development of effective – and ethical – algorithmic agents, it is heavily dominated by the voices from the computer/data science background. The existing literature on ethics currently lacks how different organisational actors define ethics in relation to their work interaction with algorithms and how they transform or challenge these work interactions. It is crucial to unearth and understand these work challenges since, firstly, there is research that underscores actors' activism, contestation, and resistances against algorithms as novel organisational dilemmas (Cameron and Rahman, 2022), and secondly, consideration of these human active behaviours as ethical conduct (Alakavuklar and Alamgir, 2018; Weiskopf and Hansen, 2023).

In line with the mentioned discussions, I adopt a Foucauldian lens and argue that there are nuances embodied in the conduct of organisational actors in line with ethics. In

doing so, the uncovered anxieties, concerns, and questions of some of the organisational actors due to the utilisation of algorithms are aligned well with the Foucauldian concept of activism and resistance (Heller, 1996). Although previous studies have identified and juxtaposed employee contestation and resistance as 'ethical activism' (Alakavuklar and Alamgir, 2018; Crane et al., 2008), algorithmic work context is a novel organisational phenomenon with only limited research on how 'ethics of resistance' is exercised. This research, however, benefits from Foucauldian notions, including subjectification and resistances, and applies them to theoretically understand how organisational actors' conducts are aligned with ethical activism. Moreover, by drawing on the underexplored Foucauldian concept of 'ethical individuals' (Skinner, 2013), I argue that although the power/knowledge of data science creates a novel working discipline through the introduction of algorithms, there are instances in the conduct of individuals that align well with Foucault's ethical subjects. As such, I showcase the nuances in discourses, statements, and behaviours of some organisational actors. By doing so, I respond to the empirical research call from Weiskopf and Hansen (2023) and surface the subtle actions of people with regards to the existence of 'space for ethics.'

1.5 The Research Context: A Criminal Justice Organisation

The researcher has managed to gain access to a criminal justice organisational setting in Europe. This organisation operates at a national level with thousands of employees divided between 12 regional districts, offering civil services including prison management, probationary practices, court hearings, and tribunals. This organisation is also highly influenced by parliamentary decision-making and political discourses. Over the past two decades, this organisation has gone through several stages of computerisation and digital transformations. Nearly a decade ago, they started introducing algorithmic (predictive) tools to enhance human-based practices. At the time of this study, the organisation was using several algorithmic predictors, including one for the assessment of risk of harm as well as a tool for predictions of serious re-offences. The algorithms used in this criminal justice organisation were developed based on actuarial science. Actuarial science is the combination of mathematics and statistics disciplines with the aim to assess risk for different purposes such as insurance and pensions. The data scientists and IT engineers at this organisation used actuarial science to design and implement their algorithms to be able to effectively predict risk(s) associated with different practices in their service. The scale and the extent of technological disruption that took place in this organisational setting were amongst the main reasons the researcher approached them for participation in this study.

The concept of ‘ethical conduct’ was raised in initial discussions with the senior management of this legal system. The leaders expressed their concerns around the potential ethical issues of using algorithmic predictors. Interestingly, the study coincided with the dawn of ChatGPT and generative AI (Yu, 2023) and the rumours that many democratic Western governments were considering banning the use of ChatGPT for the civil servants due to data privacy flags (Trendall, 2023). The surge in ethical arguments from both media and academia convinced the organisation to approve this study.

1.5 Outline of the Methodology

A social constructionist approach to epistemology, combining the findings of a qualitative case study strategy (Yin, 2018), in-depth qualitative interviews, and analysis of organisational documents, is taken in this research. This qualitative research approach was deemed the most appropriate strategy to unearth the ethical discourses as deeply as possible concerning the perspectives of key organisational actors in the criminal justice organisation of this research. The semi-structured qualitative interviews focus on the perspectives of actors around the utilisation of algorithm tools, their advantages, limitations, and any ethical issues. Incorporating organisational documents in the analysis helped to gain a comprehensive understanding of the similarities and contradictions of viewpoints from the actors and organisation itself which is the publisher of the documents. The combination of these two sources of data provided the researcher with a more detailed picture of what and how the ethical dimensions of algorithms are understood by different stakeholders.

Due to the Foucauldian theoretical perspective of this study, the research has subscribed to Foucauldian Discourse Analysis (FDA) for analysing the qualitative data. The interplay of Foucauldian theories of power and governmentality (Heller, 1996) necessitates an approach to qualitative data analysis that enables the researcher to elicit the mentioned concepts. Furthermore, the FDA method considers ‘discourses’ as power instruments/systems that shape the social world (Arribas-Ayllon and Walkerdine, 2017). Thus, it can provide the researcher with a tool to uncover how and the extent to which ethical discourses shape the work experiences and interactions of actors with algorithm agents.

1.6 Outline of the thesis

This thesis is divided into seven major chapters. Following this introduction, chapter 2 offers the reader a critical review of the existing literature around AI/algorithm ethics. Chapter two begins by providing a definition of AI and algorithms from the angle of business and organisation studies literatures, outlining their applications, advantages, and benefits of

these intelligent agents in the work context. Following this, the chapter examines the literature around the human-algorithm interactions and highlights the landscape, impacts, and emerging challenges in these interactions. In this body of literature, it is noticeable that ethics has become a momentous point and that the scholarship has raised concerns around work collaborations with the algorithms. As such, the author highlights how ethical issues can emerge and might affect those who use algorithms or are being targeted and/or processed by them. With this in mind, the chapter examines the AI/algorithm ethics literature and highlights the dominant focus on the innate technological issues, such as biased input data. The author, subsequently, underpins the overemphasis from a computer/data science angle on algorithm ethics and highlights the scarcity of contextualised and empirical research within organisational research. In that regard, the author outlines the emerging notions of power, agency, and resistance within a burgeoning branch of algorithm ethics literature that have been, to date, relatively underexplored. Consequently, the relevant literature is critically explored, and the need for a novel theoretical perspective is highlighted. Foucault's theoretical lens is introduced towards the end of chapter 2, which is aligned with the mentioned growing literature on algorithm ethics and can help to decipher ethical nuances from subjects' perspectives.

In chapter 3, I offer the Foucauldian theoretical perspective adopted to explain the findings of this research. To do so, this chapter critically explains the key concepts that are most relevant to this research. The chapter reviews the literature around Foucault's theories of discourse, power/knowledge, and subsequently, governmentality. This chapter further justifies why Foucault's theories are a way forward in relation to exploration of the ethics of algorithms. In addition, chapter 3 outlines how the mentioned Foucauldian concepts will be used in subsequent chapters to explore and interpret the findings of this study.

Chapter 4 outlines an overview of general methodological aspects and justifies the research reasoning, strategy, and design for this thesis. It explains the reasons for the choice of qualitative case study design and outlines the source of data used to carry out this research. In addition, the chapter highlights the ethical considerations of conducting this research. The chapter also explains in detail the methods utilised to analyse and interpret the data and reach the findings of chapter 5.

In chapter 5, I present the findings of this research based on the semi-structured qualitative interviews and documentation analysis. Chapter 5 discusses the findings through the established literature on ethics in order to address the research questions. This chapter provides a detailed picture of three dominant discourses found through the analysis data. First, the chapter demonstrates the influence of expertise power/knowledge (of the data

scientists) in the adoption of algorithm tools and highlights the ethical dimension within algorithmic work practice and the marginalisation of other 'voices' in the mentioned processes. Second, the chapter highlights the reactions and attitudes of different organisational stakeholders towards the algorithmic technologies. It shows that on the one hand, there is conformity and compliance amongst some of the actors in algorithmic work practices. On the other hand, there is a discursive shift amongst other organisational members, which illuminates their self-reflexivity and awareness around ethical shortcomings of algorithms. And finally, this chapter further expands on the mentioned discursive shift and explains how particular stakeholders question the practicality and ethicality of algorithms and take action to ensure the delivery of an ethical service.

In chapter 6, I will discuss the findings in relation to the Foucauldian theoretical perspective of this research. This chapter expands on the findings in relation to established literatures on AI/algorithmic ethics as well as organisation ethics and explains how the research's findings shed light on the underexplored corners of the relevant literature(s). Moreover, this chapter illustrates the theoretical and empirical contributions of this study, limitations, and recommendations for future research.

This thesis concludes with chapter 7 with some closing statements and offers a number of practical implications for algorithmically empowered work contexts.

CHAPTER 2: A Review of Literature on algorithmic work practices: benefits, challenges and ethical implications: A Foucauldian perspective.

2.1 Introduction

This chapter reviews the existing literature on algorithmic work practices with a critical look to unpack the dominant ethical debates around these technologies. It starts by defining algorithmic technologies, focusing on their practicality, perks and benefits for organisational procedures. Following this, it reviews the literature around the effectiveness of algorithms, particularly looking at the emerging debates on challenges of human-algorithm work interactions. Subsequently, the chapter examines the growing concerns in organisational scholarship regarding the ethicality of algorithmic tools. It argues that although the existing body of knowledge on algorithmic ethics has empirically uncovered many ethical issues around algorithmic technologies (Tsamados et al., 2022), it largely lacks a theoretical angle that comprehensively explains these ethical concerns. This chapter then introduces Michel Foucault's theories and concepts in order to depict a new theoretical lens that would aid in explaining and addressing the ethical/moral questions of algorithmic processes. The short introduction to Foucauldian philosophy is intended to explain and justify why Foucault's work promises a path forward for the exploration of algorithm ethical discourse. Indeed, the next chapter reviews Foucauldian literature in more detail and structures the theoretical lens of this research.

2.2 What is an Algorithm?

The dawn of fourth industrial era – 4.0 – is particularly distinguished by the rise of Artificial Intelligence (AI), Big Data, and computational sense-making (Makarius et al., 2020). Before any attempt to discuss the organisational impacts of AI/algorithms, this review explores a few key concepts and highlights the heterogeneity of the definitions around 4.0 technologies. The term '*Artificial Intelligence'* is used ubiquitously by scholars, yet there is no universal definition for it. For instance, Daugherty and Wilson (2018) picture AI systems as sophisticated computer programmes capable of automating and/or augmenting many processes, whereas Boden (2018) defines them as entities that "*seek to make computers do the sorts of things that minds can do"* (p.3). However, Jia et al. (2018) argue that AI is an interdisciplinary science with the aim to mimic human consciousness, cognition, and capabilities. It is crucial to note that AI is an umbrella term for a number of technologies including rule-based [machine learning] algorithms, natural language processing, neural networks, and deep learning. The concept of 'Big Data' is also a computer science project

compromising several elements. First, it is structured based upon massive, complex and varied datasets (Sagiroglu and Sinanc, 2013). Second, it is produced via randomly and/or purposefully selected sources, from internet searches and shared content on social media to vital signs recorded by wearable gadgets (e.g., smart watches or fitness bands). George et al. (2014) argue that 'bigness' in the term *Big Data* is no longer about size of a dataset but, rather, "*how smart it is*" (p.321). The existence of Big Data is undoubtedly an important factor in the design and development of AI technologies. The reason is that the main purpose of an AI agent is to mimic human cognition and actions. Humans learn, react and evolve through interactions with other humans and their surroundings. In terms of analogy, what human's interactions collect as information is similar to Big Data fed to AI in order to exist and evolve. Having smarter and more accurate datasets means better – more reliable – AI agents.

The term 'algorithm' – which refers to the technology implemented in this research's case study – is also difficult to define as it branches into many sub-sections within AI science. For example, Hill's (2016) definition of algorithm seems thorough but cumbersome: "*A finite, abstract, effective compound control structure, imperatively given… accomplishing a given purpose…under given provisions*" (p. 44). Recent revelations on algorithms, however, are less tedious. For instance, Lindebaum et al. (2020) consider algorithms as computer-based rules or calculations for automated decision making and/or problem-solving. Inspired by the optimisation capabilities of algorithms, Zhou et al. (2022) defines them as powerful computational systems able to make decisions based on rules and mathematical models through evaluation of available data. Algorithms are also known as 'black boxes' (Kim et al., 2020), meaning that their internal processing is opaque to the human end-users, regardless of how transparent inputs and outputs are (Geiger, 2017). The black-box characteristic of algorithms raises concerns in terms of trustworthiness, reliability and transparency of algorithmic outcomes (Durán and Jongsma, 2021).

The literature surrounding 'immersive technologies' has often pointed out the importance of AI and algorithmic tools as complementary technologies that can enhance the functionality and immersive features of such technologies (Butt et al., 2021; Orea-Giner et al., 2022; Sung et al., 2021). Immersive technologies are defined as tools or systems designed to blend the physical and virtual worlds. In other words, immersive technologies aim to make the boundaries between the physical and virtual world blurry, enabling the user to immerse him/herself in the virtual experience (Newbutt et al., 2020; Suh and Prophet, 2018; Xi et al., 2024). For instance, some immersive technologies use sensory information to synthesise a virtual environment where the user is able to interact with both physical and virtual objects (Suh and Prophet, 2018). The three highly praised immersive technologies include augmented reality (AR), virtual reality (VR), and mixed reality (MR). These

technologies can enhance the user's senses, including visual or aural, with digitalised information [which is called AR] (Wei and Yuan, 2023). VR tools blend software-generated objects with the real-world environment, enabling an interactive experience for the user (Tom Dieck and Han, 2022). MR technology is considered the continuum of virtual to real environments created by VR and AR systems (Suh and Prophet, 2018). Immersive technologies, *per se*, are not able to mimic human cognition/actions or autonomously make decisions, as opposed to AI and algorithmic technologies (Sung et al., 2021). AI and algorithms are data-driven tools employing complex procedures such as machine learning or neural networks with the aim of making predictions or solving problems (Prikshat et al., 2023). That said, AI and algorithmic tools have the potential to complement immersive technologies (Jagatheesaperumal et al., 2024). As such, there is emerging research that highlights how advanced algorithms and AI can be utilised alongside VR or AR tools to better process sensory information and engender a more interactive environment for the user (Soliman et al., 2024; Sung et al., 2021). Although it is evident that algorithmic tools can enhance the user experience of immersive systems, these two technologies are not synonymous. There is also nascent research that looks into the ethical implications of immersive technologies, highlighting issues from manipulation of virtual images and trustworthiness of the content (Sánchez Laws and Utne, 2019) to mental or psychological impacts on users (Peña-Acuña and Rubio-Alcalá, 2024). As it will be explained, there are, indeed, overlaps in terms of particular ethical dimensions between algorithmic and immersive technologies. For instance, the issue of safety and protection of the user (Southgate et al., 2019). However, AI and algorithmic technologies pose different ethical dilemmas due to their predictive nature, which will be discussed later in this chapter.

For the purpose of this research, however, and given the nature of the tools used in the case study, I will use the term 'algorithmic tools' in order to refer to the mentioned technology. Furthermore, considering the heterogeneity of concepts and the notion of *ethics* as the core discussion of this thesis, I adopt Kraemer et al.'s (2010) connotation of algorithms, illuminating them as computerised entities implemented to solve a wide range of problems. The attention on *'solving a range of problems'*, nonetheless, matches the tool(s) used in the case study, which is a predictive algorithm designed to determine the likelihood of risky conducts (Leicht-Deobald et al., 2019).

2.3 Applications of Algorithms for Organisational processes

The exciting promise of algorithmic technologies for organisational procedures has been widely disseminated in the literature, underlining their potential capabilities to streamline

many administrative procedures (Klumpp, 2018; Metcalf et al., 2019; Sutton et al., 2018). Perhaps a logical way is to categorise algorithmic applications into two sections: Algorithmic automation (Eglash et al., 2020) and algorithmic augmentation (Grønsund and Aanestad, 2020). The difference between these two draws upon the extent to which algorithmic technologies reconfigure the role of humans at work. In other words, there is a technical difference in the sense that algorithmic automation leads to delimitations on human control and enforces its *prescriptive* feature (Leicht-Deobald et al., 2019) on organisational practices. On this basis, we can tap into the organisational implications of algorithms and illustrate their merits for organisational processes.

The related stream of research in organisational studies has substantially scrutinised the application of AI/algorithms to inform administrative decision-making (Bader and Kaiser, 2019; De Laat, 2018; Duan et al., 2019; Farrokhi et al., 2020; Tambe et al., 2019), mass surveillance and monitoring purposes (De Vaujany et al., 2021; Jarrahi et al., 2021; Zuboff, 2019), supply chain and logistics management (Kosmol et al., 2019), and in the banking sector, including credit checks (Langenbucher, 2020), loans (Abuhusain, 2020) and mortgage application processing (Guler, 2015). Indeed, the practical implications of algorithmic technologies are noticeably vast, making it impossible to name all of them. As Dwivedi et al. (2021) reveal:

> "*AI technology is no longer the realm of futurologists but an integral component of the business model of many organisations and a key strategic element in the plans for many sectors of business…*" (p.2).

That said, I briefly touch upon some of the algorithm implications for organisational processes with the closest relevance to the core focus of this paper, *ethics.* On this occasion, the emergence of intelligent algorithms in Human Resource Management (HRM) has significantly elevated HR's strategic position (Vassilopoulou et al., 2022; Chowdhury et al., 2023; Gikopoulos, 2019). Algorithm-empowered HR systems are able to undertake tasks such as responding to employee enquiries, workforce scheduling, and performance measurement. Accordingly, the use of algorithms frees up time for HR practitioners to focus on non-banal and creative aspects of their jobs (Budhwar et al., 2023). Additionally, algorithmically informed HRM decision making (e.g., for selectin and recruitment, CV screening) helps to mitigate potential risks of human biases or stereotypes by ensuring inclusivity and diversity of collected data (Chowdhury et al., 2023). HR Analytics has also benefited from the algorithmic tools. Intelligent algorithms are able to mine, store, and make sense of workforce data to guide decision-making. Research identifies that integration of algorithms into HR analytics brings more precision, accuracy and flexibility into practices

(Tambe et al., 2019), with further fruitful outcomes such improved employee performance, teamwork, and knowledge sharing (Chornous and Gura, 2020). Overall, the use of algorithmic tools to inform people management decision-making gives senior leadership the leverage to better identify patterns and opportunities for enhanced knowledge production and transfer (Sestino and De Mauro, 2022).

Neo-liberalism and gig economy (Ganti, 2014) schools of thought have also played substantial roles in the introduction of algorithmic systems in organisations (Cameron and Rahman, 2022). Fairclough (2003, p.5) defines neo-liberalism as a political project with the aim to re-structure and re-scale socio-cultural relations with the limitless demands of global capitalism. Gig economy work philosophy has emerged from neo-liberalism, which includes mostly irregular work schedules that are offered and managed by digital platforms (Newlands, 2021). The technological infrastructure of many gig economy – aka *digital labour* – businesses such as Uber, JustEat, Upwork is bolstered by algorithmic agents with the aim of improving the workforce's performance via constant monitoring and evaluation. Research suggests that employee algorithmic monitoring [surveillance] maximises financial metrics by reducing the administrative workload on the gig economy-platform organisations (Bucher et al., 2021). In addition, relevant studies indicate that the outputs produced by algorithmic surveillance tools are much more *objective* than the situations where the observant is a human (Newlands, 2021; Jarrahi et al., 2021).

The applications of intelligent algorithms go beyond the mentioned administrative practices of businesses, with contributions to medical/healthcare services (Sun and Medaglia, 2019), cybersecurity (Syed, 2020), global mass surveillance (Munro, 2018; Minocher and Randall, 2020) and smart hospitality concept (Buhalis and Leung, 2018). As explained above, many businesses and organisations are now inclined to reshape their organisational strategies to better fit algorithmic agents. However, this technological shift in strategies requires a new partnership between the human workforce and intelligent machines (De Cremer and McGuire, 2022). In the next section, I explore the concept of human-algorithm collaborations at work and discuss the growing challenges of this new organisational partnership.

2.4 Human-Algorithm work collaborations: The landscape, impacts and challenges

The concept of human-algorithm work partnership refers to a situation where computational programmes support human actors in solving organisational problems (Bader and Kaiser, 2019) or assist organisational actors in their decision-making (Susse et al., 2021). Algorithmic applications such as autonomous [and swift] data mining, tracking and analysis

make them ideal candidates for both prediction and prescription in many work interventions. Additionally, situating algorithms as analytical work partners accelerates the undertaking of knowledge-based practices, which, not long ago, were known as the exclusive realm of human cognitive abilities (Jarrahi, 2018).

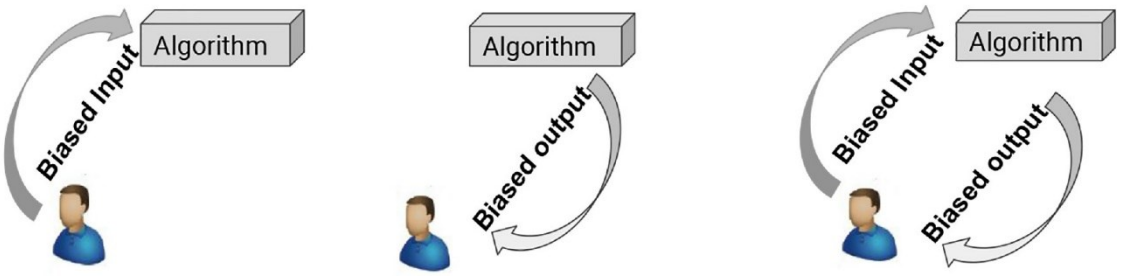
Despite the highlighted benefits in human-algorithm work collaborations, the relevant literature identifies many challenges in these novel work partnerships (Bucher et al., 2021; Jarrahi, 2018; Bader and Kaiser, 2019). Namely, De Cremer and McGuire (2022) examine employee responses to the implementation of an algorithm thinker for automation of managerial decision-making processes. Their findings suggest that the workforce's response to algorithmic work is profoundly influenced by their perception of *fairness* in the context of their work. That is to say, the human workforce perceives the employment of algorithm agents as *unfair* since they believe autonomous decision-making obliterates the key element from this process: "*humanity"* (De Cremer and McGuire, 2022). Interestingly, the participants of the study did not support the notion to remove algorithm agents, yet they reported that it is much fairer if human counterparts have the final say in human-algorithm partnerships.

Research investigates how implementation of algorithmic technologies for sales purposes unintentionally resulted in symbolic conformity of human employees, which then led to full integration of the autonomous technologies (Pachidi et al., 2021). It is argued that this unorthodox conformity was amplified by pro-algorithm stakeholders – who advocated the existence of technology, labelled as *Technologists* – and consequently symbolic conformity of those whose work practices were changed due to the rise of algorithms. Such discussions bring to light the transformative power of algorithms that limits human-users' autonomy and leads to the decay of human agency and self-determination (Floridi et al., 2018).

The literature around human-algorithm interactions is highly structured by the integration of computer science with business management scholarship, aiming to fathom the deep nuances of human-algorithm interactions (Wolf and Blomberg, 2019; Sun et al., 2020; Peeters, 2020; Tarafdar et al., 2023). In this vein, the paper by Sun et al. (2020) touches upon the volatility of bias in human-algorithm relations, highlighting it as a fluid premise that dynamically and iteratively affects this parentship. They argue that the initial data used for training and optimising algorithmic tools is produced based on human's actions in the sense that human are the responsible parties in selection, filtration, and deletion of what is fed to the algorithm (e.g., Netflix suggestion algorithm). The input data for algorithm creation can potentially be optimised based on tastes, preferences, or mindsets of a small group of people within the population, whilst the opposed undesirable feedback is removed from of the data (Leicht-Deobald et al., 2019). As a result of this iterated filtration, the

prediction, recommendations, or any decisions made by an algorithm for the human users might be essentially biased, purposive, and fallacious. The continuity of iterated bias in algorithms not only negatively affects the efficiency of algorithms – through imbalances or inequalities of the decisions – but also human-user's cognition and learning, as they are stuck in a continuum of biased *feedback loop* (Sun et al., 2020) [Figure 2.1].

Figure 2.1 The issue of feedback loop directly adopted from Sun et al., 2020.



Algorithms are capable of informing – or so-called – *shaping* people's choices. They constantly nudge, evade, constrain, and undermine human's autonomy, even if it is done unintentionally (Taddeo and Floridi, 2018). Tsamados et al. (2022) allude ongoing concerns of human-algorithm relations, arguing how it can jeopardise human's volition. It is recommended that human employees be kept '*in-the-Loop'*, meaning there should be collaboration and contribution between algorithm designers and other stakeholders (Milano et al., 2020; Tsamados et al., 2022). Peeters (2020) also taps into the issue of the decay of human discretion in algorithmic interactions, yet adds the notion of 'keeping humans in the loop' as the moot element in algorithmic applications and decision-makings. He brings attention to the *agency of algorithms* (Peeters, 2020), which is an assumption of the extent to which algorithmic technologies are capable of shaping human societies (Floridi and Taddeo, 2016). A technological artefact (e.g., an algorithm) can have politics and agency in two ways: either via deliberate design to exercise power or bias, or by the inherited power that a technological apparatus operates or functions. A notion also conceptualised by Bruno Latour in actor-network theory (Latour, 1987), in which the agency dynamics are linked to both human and non-human actors, who interact and relate to each other in particular ways, consequentially shape societies. Theoretically, this applies to the dawn of algorithmic decisions-making as it may explain the behavioural challenges in humans and algorithms interactions: On the one hand, an algorithm by design/nature is able to proliferate bias or stereotypes. On the other, algorithms have the power to influence or determine human behaviours, actions, and decisions (Curchod et al., 2020). Peeters (2020) also advocates the

concept of algorithm agency, highlighting the importance of humans' role in overseeing and/or overriding the algorithmic outcomes:

"*The problem is not only what algorithms do to people, but also what people do with algorithms"* (p. 518).

According to the mainstream literature of human-algorithm interactions, human agency is a crucial aspect, with a consensus amongst many scholars giving centrality to humans. However, most offered solutions are hypothetical, built upon the finding of many studies (e.g., Durán and Jongsma, 2021; Amitai and Oren, 2017; Taddeo and Floridi, 2018; Russell et al., 2015). This highlights a void in the existing literature in the sense that such recommendations require more research, especially that collected through human end-users. Peeters (2020) also indicates that the issue of algorithmic agency, particularly in public policy and administration literature, has received only limited scrutiny to date. Hence, I follow this strand and problematise how insistence on human agency in algorithmic partnership is constructive for the discourse ethics in algorithm.

2.4.1 People perception of algorithmic work transformations

What has been reviewed so far regarding the human-algorithm interactions predominantly considers the centrality of humans in algorithmic work practices. Yet, the existing literature features the ambiguity of algorithmic outcomes with human users' responses and merges them with the relevant ethical debates (e.g., Allen and Choudhury, 2022; Mittelstadt et al., 2016; Jago, 2019). In his experimental study investigating people's mindsets around algorithmss authenticity, Jago (2019) finds that people see algorithms as less authentic compared to themselves. His research suggests that people find it difficult to situate algorithm as authentic instruments not because they lack sincere features, but because algorithms are envisaged as tools incapable of generating *ethical* authenticity. As per the practical implications, he suggests collecting people's judgements around the authenticity of intelligent machines and using those as a framework for future algorithm development. As such, inclusion of human users in the development processes may positively contribute to bolstering the authenticity of algorithm agents. The participatory approach to algorithm design and adoption has been identified and endorsed by other scholars which can positively influence addressing ethical challenges of algorithms (Leicht-Deobald et al., 2019; Suseno et al., 2021). Whilst such studies within organisational literature envisage novel research avenues to human-algorithm work interactions, they simultaneously underscore how pivotal employee perception is with regards to ethical dimensions.

Further research on humans' perception of algorithm decision-makers has found a dyadic role within these systems: On the one hand, algorithms present themselves with user-friendly interfaces that sometimes increases humans' involvement in decision-making. On the other, these tools are able to subtly dictate decision to humans which may lead to detaching humans from decision-making (Bader and Kaiser, 2019). This dual behaviour may create a dubious situation in organisational environments due to disproportionate and/or imbalanced decision making and consequently may lead to workarounds and deferred decisions. Empirical findings suggest that algorithmic obscure functionalities (Safdar et al., 2020), or their Black Box nature (Adadi and Berrada, 2018) may result in human end-users questioning their cognitive abilities and either over-relying on algorithmic outputs or detaching themselves from it. Their argument indeed overlaps and supports with the earlier debate on how algorithm end-users perceive their work interactions with their intelligent partners. It also raises a caveat on how employees perceive and situate ethical values in their work partnership with algorithms; a profoundly overlooked point in relevant studies (Ouchchy et al., 2020).

The existing literature has touched upon the issue of *algorithm aversion* (Dietvorst et al., 2015) as a catalyst for ineffective human-algorithm interactions: a staple component also found in the relevant ethical algorithm literature (Walker et al., 2021; Weiskopf and Hansen, 2023). Algorithm aversion refers to a situation where the human end-user refuses to utilise or accept the algorithm's outputs or forecasts. Discussions around human reluctance to use algorithms are not entirely new. Dietvorst et al. (2015) used this term in their work, and many scholars have subsequently looked into the issue of algorithm aversion with the intention of understanding the antecedents of this behavioural phenomenon (e.g., Allen and Choudhury, 2022; Burton et al., 2020; Logg et al., 2019). Dietvorst et al. (2015) highlight two main reasons that cause algorithm aversion: First, frequent errors by algorithmic tools make human users hesitant to put trust in the generated forecasts. Second, human users expect algorithms to outperform them. Yet, as algorithm forecasters are less likely to reveal a '*prefect*' prediction compared to those generated by humans, this makes humans sceptical about relying on algorithm outputs.

An empirical study by Allen and Choudhury (2022) explores the level of algorithm acceptance or aversion in relation to the domain of work and finds that algorithmic work enhancement is linked to employee work experiences. Thus, whilst a low-experienced workforce showed more willingness to embrace algorithmic tools, this level of acceptance shifts to *aversion* (Dietvorst et al., 2015) amongst those employees with medium to high level of work experience. To mitigate such irregularities, in particular for those with higher work experiences, Dietvorst et al. (2015) recommend that designing algorithms should be

transparent and inclusive. Transparency and inclusion in the algorithms design processes enable employee end-users to be part of the algorithmic work culture and exercise more control in such hybrid environments.

Although research around algorithm aversion stipulates essential actions for an effective – and ethical – adoption and implementation of algorithm agents (Burton et al., 2020), there is only minimal support for the raised the questions on employees' understanding of algorithm ethics (Charlwood and Guenole, 2022) which is a pivotal tenet in algorithm scholarship. The above discussions of the literature surrounding the premise of human-algorithm interactions underscored that the efficiency of the relationship algorithmic intelligent agents and humans very much depends on the people's perception of algorithm's genuineness, authenticity and ethicality (Holford, 2019; Jago, 2019; Dietvorst et al., 2015). Despite the unfolded factors affecting human-algorithm relations and the means to mitigate algorithm mistrust, scholars still argue that there are many unknown nuances in the concept of human-algorithm engagement (Van Esch and Black, 2019; Seeber et al., 2020). Indeed, the literature on human-robot collaboration has made attempts to identify and signpost the antecedents for boosting this relation (Simon et al., 2020; Schniter et al., 2020; Sherwani et al., 2020). Yet, it is worth noting that not all robotic technologies are endowed with intelligent reasoning and/or autonomous decision-making capabilities. Hence, robotic apparatuses establish a different interactions category in relation to human end-users, outside the scope of this research. Even considering the overlaps in the literatures of human-algorithm/ robotics, it is still ambiguous whether the emerging ethical issues can shift or inflict damage on the human user's mindset about algorithmic tools. Tsai et al.'s (2022) literature review raises questions around *robotics* ethical considerations in terms of personnel replacements, trustworthiness of robotics and their subjective decision-making, whilst criticising existing research's limited focus, which has already ignored the intelligent algorithms. Broadly speaking, it seems the literature around human-algorithm collaborations lacks a theoretical perspective around the ethical side of these interactions. Said differently, the relevant literature calls for further research on the impacts of ethics in human-algorithm work partnerships (Baum, 2020; Tsai et al., 2022; Tursunbayeva et al., 2022). Addressing this gap will contribute to the literature by providing a comprehensive understanding of the existing ethical factors that affect human-algorithm relations from the perspective of the workforce. Additionally, this research will have implications for human-robot collaboration literature by providing empirical evidence on whether and how ethical issues impact the actors' work relationships in the hybrid human-robot work milieu.

2.5 Growing concerns around the ethical issues of algorithms

A significant proportion of literature tends to postulate that adoption and implementation of algorithm and their partnership with workforce agencies have ethical implications within organisations (Ajunwa, 2020) and societies (Murray and Flyverbom, 2020). Yet, as it will be shown, little research has been done to explore the employees' viewpoints on the ethical issues in algorithmic tools and what the discourse of ethics is amongst the human end-users.

The concept of ethics in algorithm technologies has a wide range of definitions found in many relevant disciplines from data science and informatics (Mittelstadt et al., 2016) to psychology (Bigman et al., 2022) and organisational studies (Vassilopoulou et al., 2022). Although it seems defining ethics of algorithms is more a matter of perspective, it is predominantly mapped out as any evidence, effects, outcomes and/or implications that might go against the ethical ethos or principles of human societies. In other words, ethical algorithms are essentially those whose existence and impacts are in accordance with wider societal ethical norms and values (Kraemer et al., 2010).

Studies exploring the ethical aspects of algorithms – mainly through the computer and data science disciplines – raise awareness that algorithms are not always ethically neutral or innocuous. For instance, Kraemer et al. (2010) argue that there are certain moral issues with algorithms which particularly stem from their *value-laden nature*. Value-laden algorithms are designed and shaped by subjective variables, parameters, opinions, or inputs, determined by the designers (Martin, 2022). As Kraemer et al. (2010) discuss the value-laden characteristic of algorithms as an ethical issue, as it makes algorithmic judgments subject to optimisation and/or manipulation by design teams. Kraemer et al. (2010) invoke *Kantian ethical discourses* (Wood, 1991) and recommend that identification and justification of ethical values in algorithms can be assigned to the users. A user-centred approach to the identification and justification of ethics is a pivotal concept for this research, of which the existing literature shows degrees of scarcity. The literature underpins some key issues around the ethical dimensions of algorithmic tools, calling for more research on how these technological tools can be designed and governed in a *"socially good"* manner (Mittelstadt et al., 2016; Tsamados et al., 2022). Research highlights how pivotal it is to contemplate what the ethical risk are, what the *potential for goodness* is, and how socially good outcomes can be achieved by algorithmic technologies. However, as scholars emphasise the role of human agency, actionable conduct, and supervision, little is known around how the key organisational actors understand and debate algorithm ethics within the context of their work practices. In the following, I will tap into this scarcity within the algorithm ethic literature and narrow it down to algorithmic decision-making.

2.5.1 The Ethical Issues in Algorithmic decision making

The existing literature has reported that the emergence of state-of-the-art technologies such as Artificial Intelligence, Big Data and Internet of Things will transform the dynamics of many organisations in the coming years (Chowdhury et al., 2023). Although the research around the merits and benefits of utilising algorithms is expanding exponentially (e.g., Zheng et al., 2017; Jöhnk et al., 2020; Sherwani et al., 2020; Guan et al., 2020), there is now significant emphasis on the moral and ethical dimensions of automated/augmented decisions and procedures.

Critical/emancipatory studies, in particular, have aimed to demonstrate the embedded ethical issues of algorithms within a variety of subjects and disciplines. As such, there is an attempt in that literature to signpost and bolster the awareness of biases, stereotypes, and discriminatory outputs of algorithms (Mittelstadt et al., 2016; Tambe et al., 2019; Amitai and Oren, 2017; Baum, 2020). Manifestation of inequalities and injustices are found particularly in algorithmic HR recruitment decision-making processes (Mujtaba and Mahapatra, 2019). Instances of such inequalities were found in people management contexts where, for example, women's CVs were eschewed by an algorithm recruiter (Bigman et al., 2022), or certain privileges or promotions were given disproportionately to male candidates based on algorithm's predictions (Tambe et al., 2019). Algorithmic tools have also been found marginalising or making unjustified differentiation against non-binary and transgender employees. A study by Vassilopoulou et al. (2022) deeply focuses on the biases of algorithmically guided decision-making in HRM practices. In that respect, the paper argues that the algorithmic decision-making in HRM is directed by scientific orientations and orthodoxies, "*giving unfounded prioritisation of scientific methods over and above the moral values and reasoned arguments"* (p.316). Scientific and positivists philosophies in algorithmic decision-making may deprive human users of their agency and suppress them from questioning rules and regulations (Vassilopoulou et al., 2022). This is imperative considering that some organisational procedures – such as HRM – necessitates degrees of human agency and autonomy (Caldwell et al., 2010), whilst the introduction of algorithmic decision-making limits and circumvents practitioners' volition and virtue (Mennicken and Miller, 2014). In a similar vein, a paper by Leicht-Deobald et al. (2019) argues that algorithm decision supporters are neither entirely objective nor ethically neutral. To explain this issue, the authors discuss that the implementation of algorithmic systems for administrative decisions making can negatively affect employee's sense-making and amplify organisational conformity, whilst resulting in blind trust in external rules and regulations. Similarly, Charlwood and Guenole (2022) study about HR algorithmic automation warns us of the

threats that algorithms might pose on organisational procedures such as perpetuating systemic bias, unfairness and other dystopian consequences.

It is suggested that integration of more concrete data in algorithms would bring further foresight into automated decision making, not only making tools more reliable, but also streamlining employees' acceptance of technology (Alaimo and Kallinikos, 2021; Fuchs et al., 2014; Günther et al., 2017; Roßmann et al., 2018). However, due to numerous ethical challenges posed by algorithm outcomes, scholarship calls for further investigation into what kind of data is best labelled as 'concrete', and how this type of data can constitute ethical algorithmic decision-making (Oswald et al., 2020).

The literature, however, offers two major solutions to rectify the ethical issues: First, the scholarship acknowledges and values the employees' ethical awareness, encouraging them to jointly engage in the ethical discourses of algorithmic decision-making (Leicht-Deobald et al., 2019). Second, studies advocate the idea of *participatory design* in which the future users of an algorithm become the co-designers, putting their ideas and values into the design (Charlwood and Guenole, 2022). Whilst such discussions in the literature have paved the way to make algorithmic decision making more ethical, the suggestions seem prescriptive and hypothetical, only taking into consideration the external elements [such as big data] as a panacea to resolve ethical concerns. It seems further research is needed to unearth the ethical discourses amongst employees to cushion ethical algorithms. Especially taking into consideration the research call by Charlwood and Guenole (2022, p.738) to carry out more *"qualitative phenomenological research… that can provide the basis for novel theoretical insights*", highlights the need for more evidence-based research on this matter. Despite the existing ethics literature often paying tribute to the role humans play in algorithmic transformations and criticising one-sided perspective of ethics, an absence is noticeable around how the ethical issues are understood, made sense and impact human users' behaviour and work experiences. In this respect, Charlwood and Guenole (2022) paper considers the ethical perspective around algorithmic HR, people management practices, outlining: *"How do they [the autonomous technologies] reconfigure division of labour and social relations within organisations? … How much agency do they [HR practitioners] have to moderate the impact of AI on their work?"* (p.738).

This section underpins that algorithm ethics literature is predominantly constructed by the theoretical viewpoints of scholarship, and lacks the perspective of organisational actors, especially the frontline staff who are situated in the heart of algorithm work contexts. Pursuing this research agenda will make contributions to existing literature on algorithm ethics by providing a more contextualised perspective of the mindsets, perceptions and

attitudes of the human users regarding the ethics of algorithm utilisation. The factual – empirical discussions and work experiences of the organisational actors will not only inform and extend the existing hypothetical principles of algorithmic ethics, but also inspire further research on how workforce agencies voice their concerns in shaping algorithmic ethics in both critical-emancipatory organisational and computer science literature. Additionally, this research contributes to existing literature on algorithmic risk prediction policies in criminal justice settings by offering an in-depth exploration of how and to what extent practitioners fathom, discuss and incorporate ethical considerations in using algorithmic assessment across a variety of interventions and rehabilitation decisions. This study seeks to construct an understanding based on the lived work experiences, discourses and actions of human users that signify the ethical values of algorithmic decision making.

2.6 Mitigating ethical concerns in algorithms: Transparency, Accountability and Responsibility

As it was mentioned earlier, the frameworks for ethical AI or ethical algorithm technologies have been showcased within variety of disciplines: from Psychology and Law to Journalism and Organisational studies. The ethical frameworks predominantly intend to elevate the ethicality of algorithms through concepts such as algorithm transparency (Robinson, 2020), accountability (Ananny and Crawford, 2018), and responsibility (Dwivedi et al., 2021). The premise of algorithmic accountability has been conceptualised around how the emerging ethical issues (e.g., discriminations, biases and stereotypes) can be governed, controlled and mitigated through human inclusion (Shah, 2018). In terms of accountability of algorithms, Kroll et al.'s (2017) revelations published in a Law outlet praise the rigorous governance of algorithmic decision-making, highlighting the emergence of unfair, unjust and incorrect computational outcomes. To ensure accountable and transparent algorithmic decision-making, they have advocated and invoked the concept of '*procedural regularity*. Procedural regularity for algorithm processes incorporates *human's supervision* (Lepri et al., 2017) of those processes to the extent that humans are relatively aware of rules used to generate an automated decision. Furthermore, it is discussed that through procedural regularity there will be consistent incorporation of standards and policies, as automated decisions are verified and justified by not just the tech/data scientists, but also by other stakeholders. Whilst procedural regularity provides sound justifications to challenges of accountability, arguments indicate it is not enough: in-depth collaboration between data scientists and other stakeholders in the context was suggested throughout the design stage in order to ensure accountability. Scholars underline that designer-user collaborations should firstly acknowledge human oversights, negligence and cognitive chasms before holding

algorithm agents accountable for their ethical misconducts (Ananny and Crawford, 2018; Hoffmann et al., 2018).

In a similar manner, studies by Diakopoulos (2015), De Laat (2018) and Janssen and Kuk (2016) have supported the idea that collaborations between algorithm designers and the human end-users may provide a path forward to develop more transparent ethical algorithms. Diakopoulos (2015) research demonstrates that the power of algorithms to undertake autonomous decision-making can result in many ethical consequences. Through qualitative interviewing the study suggests that journalists who scrutinise algorithm accountability need to achieve a computational comprehension and engage in synergetic dialogues between those with data/tech expertise. De Laat's (2018) research has also tapped into the notion of algorithmic decisions-making and its accountability from an "*overall moral perspective".* He advocates transparent design to ensure accountability of algorithms. Yet a caveat is raised in the sense that full public transparency of algorithms may jeopardise data privacy and undermine the algorithm's efficiency. The prerogative to carry out transparency evaluations should exclusively be given to auditing bodies who are able to discern and disclose relevant information according to their conventions. Meanwhile, it is argued that affected individuals or human end-users should also be able to interpret algorithmic decisions, and this interpretability can be mediated by the oversight bodies' assistance (De Laat, 2018). Janssen and Kuk (2016) editorial piece signals a warning that the emergence of algorithmic practices may lead to *technocratic governance* or a *technological singularity* if not held accountable. Technocratic governance is depicted as a hypothetical structure in which political and social institutions have lost their power to algorithmic supremacy (Janssen and Kuk, 2016). The study also underpins the difficulties of algorithmic accountability with respect to their subtlety and invisibility, making it difficult for the public to fathom an algorithm's impartiality. Nevertheless, there are encouragements and calls to carry out research for better understanding the effects and risks of data-driven algorithms, and the methods to ensure algorithmic accountability and responsibility (Dwivedi et al., 2021; Neyland, 2016).

Synthesising ethical frameworks to address ethical issues of algorithms and AI technologies has also been discussed by medical ethicists in the healthcare literature (e.g., Durán and Jongsma, 2021; Geis et al., 2019). A paper by Geis et al. (2019) unfolds the benefits of utilising algorithm technologies for radiologists, entailing augmented predication and decision-making about the patients. In addition, there are praises that algorithm-based machines enable practitioners to elevate data concentration, categorisation and evaluation. Yet, they have called for vigorous algorithmic transparency pointing out that the application of algorithms may pose systemic risks such as the issue of data stewardship, reliability of

algorithmic decisions and detection of autonomous errors. Similarly, a viewpoint is articulated by Durán and Jongsma (2021) around the ethical issues in algorithmic medical practices. In their paper they have raised the notion of *black box* nature, illustrating that some – *if not all* – internal processes of algorithms are ostensibly *opaque* [uninterpretable] for human end-users. They have discussed that adding more transparency to the opaqueness of algorithms does not necessarily rectify the emerging ethical concerns. They suggest that designing specific predictors to make algorithms more interpretable – and transparent – only helps to establish further trust in algorithmic agents, and not resolve the issue of opaqueness. The barrier to this is claimed to be within the limitation of human end-user's cognitions and information processing. To their mind, the concept of *Computational Reliabilism* (Wheeler, 2020) rectifies this challenge through admitting the human cognitive limitations. Although the study offers peripheral reference around keeping humans in the loop, little is known on how medical practitioners' acknowledgment of their cognitive limitation would advance the discourse of algorithm ethics in their work contexts.

Indeed, the medical ethics literature offers concrete arguments around algorithmic transparency and accountability. Yet the prioritisation of ethics is directed in a way to minimise the risks towards the patients, and not the medical staff. It is of course a sensible research route given the sensitivity of patient's data, methods of treatments, and their informed consent. However, a key stakeholder in this context are the healthcare professionals whose working lives are altered due to autonomous medical practices, and not much is known of their perception of algorithmic healthcare practices.

2.7 The Discourse of Algorithm Ethics

As it was argued earlier, the introduction of algorithmic technologies for organisational procedures has proliferated in recent years in line with the emergence of big data and advancements in producing data clusters (Tambe et al., 2019; Sutton et al., 2018; Jarrahi, 2018). Having said this, a key viewpoint underlying the literature in algorithm ethics is around the discourses and decision-making processes to adopt and implement algorithmic technologies (Alsheibani et al., 2018; Desouza et al., 2020; Sheehan et al., 2020). This branch in algorithm literature aims to unearth the organisational factors, events, and voices that influence and drive the strategic decisions around the adoption of algorithm technologies (Sheehan et al., 2020; Coombs, 2020). However, it is worth reviewing the relevant literature on a broader aspect of technology adoption decision-making [and aside from AI and algorithm technologies] since the algorithm adoption/implementation literature has emerged from this strand.

Since the emergence of information-based apparatuses, many scholars have aimed to identify the antecedents for organisations to adopt these technologies and ascertain how decisions are made to implement them (e.g., Langley and Truax, 1994; Zhu and Weyant, 2003; Zorn et al., 2011; Chan and Ngai, 2007; Spencer et al., 2012). For instance, Langley and Truax's (1994) longitudinal study explores the adoption processes of microcomputer drawing tools in Canadian manufacturing firms. The paper indicates that technology adoption decisions are solidified by three sub-processes: the strategic commitment from managers, the choice of technological process, and the justification to fund the technological transformation. Additionally, the study underpins that regardless of the size of firms, the decision-making practices to introduce new technologies are chiefly centralised around CEOs, top executives, or functional managers who are able to dominantly voice their needs and politically/financially justify that *need* to introduce novel technologies (Langley and Truax, 1994). In contrast, Au et al. (2003) theoretically assume that the IT adopters are very much influenced by the competition in the marketplace. The paper theoretically sets a number of propositions to support an argument that the IT adopters observe, enact, and adjust their adoption behaviours according to other rivals within the business environments. The paper consequently stipulates that IT adoption decision-makers should carefully consider the expectations of economic models within the context of their business and make consistent IT adoption decisions that would meet the desired outcomes of the marketplace (Au et al., 2003). Throughout the study, the authors talk about the decision-makers of IT adoption, yet they predominantly refer to information system [IS] managers and elites as the main decision-makers, hardly taking into consideration the perspectives of other organisational stakeholders or the ethical dimensions of IT adoption.

Chan and Ngai's (2007) study highlights that the adoption of IT-based technologies in organisations is profoundly affected by the perceived costs and benefits of such technologies, the organisational readiness, and external pressure from the competitors. The study is based on a qualitative enquiry of 10 organisations that adopted web-based training systems in Hong Kong. Although the study emphasises the level of IT knowledge amongst both top management and staff members to elevate organisational readiness, it seems the top management's opinions are prioritised when it comes to decision-making processes for IT adoption. Such arguments highlight that organisational leadership approaches are amongst the prevalent factors that influence the IT adoption decisions in organisations (Spencer et al., 2012). According to Spencer et al. (2012), a leadership perspective is a distinctive [and influential] factor in technology adoption decision-making – specifically in smaller firms – since they are defined as a "*catalyst for strategic change".* This is indeed in line with earlier discussions by Langley and Truax (1994) and Chan and Ngai (2007),

underlining the role of top management in driving the decision-making for IT/technology adoption. Latest studies on the adoption of technology in organisations also yield similar notions. For instance, a study by Falwadiya and Dhingra (2022) about the integration of blockchain technologies in government organisational settings underscores that the adoption of technology is heavily motivated by the top management's expectations to boost performance and streamline the processes. In other words, top executives as the key decision-makers decide on the adoption of a technology based on its perceived usefulness, relative advantages, and ease of use by humans in order to improve work efficiency (Falwadiya and Dhingra, 2022). In a similar manner, Aapaper by Zorn et al. (2011) explores the factors that impact the adoption decisions of ICT systems in New Zealand non-profit organisations. By adopting and advancing the institutional theory, the study demonstrates that the decisions to implement ICT are mainly catalysed by the '*self-perceived'* leaders who possess the IT knowledge and competencies and are able to scan the competitive environment and choose the ideal technological resources.

Ultimately, the review of the literature around decision-making for technology adoption indicates that the ethical implications of a technological marvel are hardly considered as concerns in the preliminary stages of adoption. However, scholars argue that understanding the ethical implications of technologies is crucial for organisational stakeholders and people when it comes to the adoption of innovative technologies (Ratten, 2012; Abara and Singh, 1993). Furthermore, the constructed theories showcased in the existing technology adoption literature highlight a limitation with regards to ethical implications. That is to say, although the majority of theories used to explore technology adoption offer momentous empirical insights on the process of adoption, they substantially overlook the ethical dimensions of emerging technologies (Pimentel et al., 1992).

In order to strengthen the argument further, I review the existing literature around the adoption and implementation of AI and/or algorithm technologies. Reviewing the AI/algorithm adoption literature is informative for this research as it is situated within the broader literature of technology adoption, yet indicates similar theoretical limitations around the ethical implications. To explain the decision-making processes that lead to the adoption of algorithm/AI tools, scholars have used a number of models, frameworks, and theories (Alsheibani et al., 2018; Sestino and De Mauro, 2022). However, amongst all the developed theories, three of them seem to be more prevalent in the relevant literature: Technology Acceptance Model [TAM] (Davis, 1989), Technology, Organisation, Environment model [TOE] (Tornatzky et al., 1990), and Diffusion of Innovation [DOI] (Rogers, 2003). In the following, I will review some of the prominent papers that have used and advanced the mentioned frameworks to scrutinise the algorithm adoption decision-making.

Pillai and Sivathanu's (2020) paper employs the TOE framework to explore the adoption of AI systems for talent acquisition in India. The study unfolds a number of antecedents that influence the decision-making processes to deploy AI technologies, such as competitive advantages of AI, market pressure, HR readiness and top executives' support. Although the talent acquisition AI is supposedly being used by many HR practitioners across their practices, it is underlined that top HR management support is pivotal in this endeavour. The support from top managers is placed at the pinnacle of importance compared to other factors in the TOE model, such as environmental [e.g., market competition] or technological [e.g., security and privacy concerns] (Pillai and Sivathanu, 2020). As literature explains, a firm's decision-making efficiency to introduce AI/algorithm technologies is believed to be enhanced significantly if a managerial adaption mechanism embraces the technological novelties. A study by Chen et al. (2021) alludes to this point by integrating the DOI theory with TOE to explain the fundamental factors that influence the successful adoption of AI technologies in the Chinese telecom industry. The DOI theory does indeed elaborate on the role of contextual factors that affect technology/innovation adoption decision-making, such as the socio-economic characteristics of organisations as well as the characteristics of the technology itself. Yet as scholars argue, DOI theory is deemed less comprehensive compared to TOE (Oliveira and Martins, 2011) and lacks the theoretical lens to investigate all organisational factors, including managerial propensity for AI/algorithm adoption (Chen et al., 2021).

The Technology Acceptance Model [TAM] fundamentally touches upon the key elements that affect the *intention* and *functionality* of technology in the adoption decision-making (Davis, 1989). With this in mind, Chatterjee et al.'s (2021) research devises an understanding of the adoption procedures of AI technologies in manufacturing and production organisations in India by extending the TAM framework. The study illustrates the importance of organisational readiness for successful deployment of AI technology, which encompasses positive influences, such as how organisational actors perceive the usefulness of AI technologies and how compatible organisational resources are to cement the AI systems. In particular, the study highlights that organisational readiness is dependent on the employees' knowledge, skills, and competencies, which means that the existence of highly trained employees not only would boost their perception of AI usefulness but will also streamline the implementation of AI tools (Chatterjee et al., 2021). In addition, the study statistically depicts the impact of strong or weak leadership support on the employees' perception of AI usefulness. Although the study exhaustively values employees' readiness in the AI adoption processes, it also highlights the managerial role on how they *should* appropriately upskill and train their employees with relevant AI knowledge and expertise.

This argument is in line with earlier studies (Pillai and Sivathanu, 2020; Chatterjee et al., 2018), which theoretically situate top management at the zenith and sole voice of AI adoption strategic decision-making.

As it was demonstrated in this section, the core factors of the theoretical framework used to explore technology adoption decision-making [including the adoption of AI/algorithm technologies] are predominantly related to the availability of organisational/technological resources and readiness as well as the socio-environmental conditions of the market. TOE, TAM, and DOI theoretical models have all made substantial empirical contributions to the literature for better understanding of the key influential factors in AI/algorithm adoption decision-making, yet they have not been exempted from the scholarship's criticism (Gangwar et al., 2014). Although such theories provide thorough roadmaps for technology adoption decision-making processes [e.g., introducing AI/algorithm tools], they often lack ethical perspectives of technological marvels. Indeed, there are some considerations in the Technology dimension of the TOE model that overlap with the ethical side of technology and reveal security and privacy concerns (Pillai and Sivathanu, 2020). Yet, such theoretical frameworks have been integrated into research projects to quantitatively analyse the relationship between ethical dilemmas and the performance and effectiveness of technology (e.g., Reinares-Lara et al., 2018). In addition, the prominent theoretical frameworks, such as TOE or TAM offer no or only limited critical arguments around the ethical nuances of technology adoption, let alone providing insights on how to identify and mitigate the concerns.

2.8 A Foucauldian theoretical lens on the discourses of ethics in algorithmic work practices

The algorithm/AI literature on the whole conceptualises ethics as a *static* issue. An issue that demands mitigation and perfection through prescriptions from the technological elites (Floridi and Taddeo, 2016). However, as discussed above, the ethical concerns of algorithmic agents go beyond the technological issues [e.g., biases in data] and do very much resonate around the perception of human end-users, employees, or those whose working lives are affected by the existence of algorithm agents (Curchod et al., 2020). I invoke Foucault's theory of *critical* discourse (Foucault, 1972), power-knowledge (Foucault, 1977), subjectification (Heller, 1996), activism and resistance (Caldwell, 2007) to further explore and unpack the ethical dimensions of algorithms from an employee perspective.

In recent years, scholars have subscribed to the post-structuralism and post-Marxism/Weberism philosophies to investigate the emergence and impacts of algorithmic/AI

work milieux in organisational studies (Walker et al., 2021; Munro, 2018; Königs, 2020). On this note, Lange et al. (2019) investigate and explain the influence of algorithms on organisational politics through Michel Serres's theory of *quasi-objects/subjects*. Newlands (2021) explores the novel algorithmic methods of control and surveillance regimes within gig economy businesses by applying Henri Lefebvre's *spatial triad* theory. Salter (2019) argues how algorithmic technologies can assist human users to avoid oversimplifications and prejudgments of power relations, with the help of Bruno Latour's *actor-network* theory.

Whilst the highlighted post-structuralist theories offer unprecedented insights, specifically to unlock the mysterious premise of algorithmic power and surveillance, they manifest degrees of indeterminacy around dominant social interventions, discourses, and rituals. Foucault's work, on the other hand, although shaped in the heart of post-Marxism/postmodernism, rejects all these labels and reflects his views on societal multiple truths and/or discourses and his quest for emancipation (Caldwell, 2007). Foucault's philosophy is aligned with constructionist epistemology, which has founded the baseline for his theory of discourse throughout his revelations as a historian/philosopher.

To date, many scholars have utilised and advanced Foucault's concepts for exploration of algorithmic work practices. For instance, Walker et al. (2021) research particularly subscribes to Foucault's notion of *Biopower,* underpinning how algorithmic tools have become a novel power technique with the aim to regulate the workforce. Weiskopf and Hansen (2022) position algorithmically driven decision-making as an improved base of organisational knowledge, arguing how algorithms limit the human workforce's reflexivity and reasoning. The study juxtaposes Foucault's notion of *Governmentality* with the emerging modality of algorithmic authoritarianism in organisations, highlighting the extent to which algorithmic processes shape a dynamic space of ethics (Weiskopf and Hansen, 2022). Although the mentioned studies have used and expanded Foucauldian work in algorithm literature, they have offered only limited theoretical insight for understanding the ethical debates of the novel tools. This research, however, utilises Foucauldian theories to particularly investigate the ethical dimensions of algorithms, developing a Foucauldian theoretical lens that can profoundly and vividly unpack the algorithms' ethical issues through the eyes of subjects.

The choice of Foucault's analyses for this research seems to positively contribute to literature since Foucault, during his scholarly work, chiefly centralised the administration and organisation of lives (Mennicken and Miller, 2014) in his work. His emphasis on human lives links to our earlier discussions of how algorithmic tools gain control of and overshadow human agency and autonomy. In order to devise the theoretical framework, I break down

Foucault's philosophical work into three themes and subsequently approach and analyse the issue of algorithmic ethics through each Foucauldian conceptual theme. On this basis, the breakdown of Foucault's work would be first, the theory of *[critical] discourse*; second, the notions of *power-knowledge* and *subjectification*; and finally, Foucauldian *ethics*, *activism,* and *resistances*. This section offers a short introduction to this study's theoretical lens with the intention of constituting how Foucauldian philosophical standpoints can be a path forward. Indeed, the next chapter provides a more detailed explanation of Foucault's analyses, including his work around *Governmentality, Discourse, Care for the self and Ethics,* accounting for the emerged critique.

2.8.1 Foucault's premise of Discourse

The linguistics viewpoint on the concept of discourse fundamentally differs from what Foucault has theorised in his book *The Archaeology of Knowledge* (Foucault, 1972). Whilst the linguistics approach sees discourse as a formal usage of language in social functions, interactions, or natural occurrences, Foucault thinks of discourses as more embedded in disciplines and material *practices* (Clegg, 1998). He portrays discipline as both the *academic* disciplines, such as medicine, sociology and psychiatry as well as disciplinary *social control institutions,* including prison, school, hospital, and so on (McHoul and Grace, 1998). According to Foucault, most of the theoretical and empirical work on discourse has a narrow focus on techniques and strategies of utterance production and recognition, known as *enunciations.* Foucault discusses that such a limit focus only scratches the surface of the meanings within the spoken or written language. Instead, he argues that discourses chiefly function through *power relations and* suggests that the analysis of discourses should traverse through the lens of power, not because a given discourse is an instrument to exercise power, but because that given discourse is a *modality* of power (Hindess, 1996). Linking Foucault's theory of discourse to the context of algorithm work commodities, one can distinguish how particular discourses – stemming from data science/engineering – subtly support and navigate the implementation of algorithmic tools and augmented decision-making practices (De Vaujany et al., 2021). A Foucauldian lens of critical discourse provides a distinct approach towards understanding the underlying reasons for the adoption of algorithmic tools as well as the functioning power-knowledge discourses that characterise algorithmic work cultures. For instance, a statement such as "*the ability for AI to overcome some of the computationally intensive, intellectual and perhaps even creative limitation of humans, opens up new application domains within [….] with resulting impacts on productivity and performance"* (Dwivedi et al., 2021, p. 2) may indicate a distinctive power-knowledge

discourse that promises an organisational supremacy via algorithmic agents while clandestinely dominating human subjects.

The review of literature identifying the ethical side of algorithmic technologies also shows an *intensification* of power-knowledge discourse (Hardy and Thomas, 2014) amongst ethicists scholarship that categorises, classifies, and objectifies human bodies. On the one hand, a branch of algorithm ethics literature highlights the issue of data impurities and imbalances, or more specifically, human biases and stereotypes that are fed back to algorithmic decision-making (Charlwood and Guenole, 2022). On the other hand, another strand of algorithmic ethics literature touches upon the manifestations of inequalities and discriminations by algorithmic predicating systems (McKay, 2020), emphasising the importance of algorithmic *fairness* through accountability and transparency (Redden, 2018). Such arguments give us a broad landscape that discourse of power dynamically circulates through the social body, yet signifies a dominant [and asymmetrical] power construct on work to make particular subjects 'known', whilst making others subordinated (Clegg, 1989). In the context of our research, ethical discourses are embedded in power relations, shaping and structuring forms of ethical regulations whilst normalising the conduct of human end-users. Yet, as illustrated above, the relevant literature presents many ethical frameworks, envisaged by data science elites without theoretically acknowledging the ethics. In this research, I argue that the Foucauldian lens of power discourses can enrich the literature by unfolding the ongoing ethic-power discourses of human employees; to hear the voices of those employees who are constituted as *objects of knowledge* (Townley, 1993) by algorithmic work regimes and reciprocally sustain the power discourses via their regulated conducts.

2.8.2 Foucault's Power-Knowledge and Subjectification

Foucault's visionary work *Discipline and Punish* (Foucault, 1977) predominantly taps into the programmes, practices, and rationalities of punitive/judicial systems that aim to regulate and normalise individuals in a sense that the subjects are self-governed and self-disciplined without exertion of coercive power. Foucault problematised power discourse not as a concept with polarity of subject and object, but as an intrinsic phenomenon that takes both forms of subjectification and objectification that make human lives *subjects* (Knights, 2002). As discussed earlier, algorithmic decision-making may pose many organisational threats, such as employees' over-reliance on algorithmic tools, leading to the decay of human centrality and agency as a result of the over-reliance. Furthermore, I discussed that a rationality for implementation algorithm predictors is to obliterate human biases and make

fair/just decisions. Yet, as I argued earlier, algorithmic tools occasionally perpetuate biases, discrimination, and stereotypes, while human end-users are subjugated and subordinated to their digital/computational capabilities. To explain this, Foucault conceptualises power relations not from the angle of top-down power hierarchies of the state or senior management, but of the actions/conducts *on* the actions of others (Raffnsøe et al., 2019) in a sense that subjects become distinctive pillars in power relations and their actions/conformities [e.g., to algorithmic outputs] further constitute them as regulated bodies.

Foucault also talks exhaustively about Bentham's concept of panoptic control (Foucault, 1977; Knights, 2002), in which subjects are under continuous monitoring practices. He explains how that *Panoptic gaze* or *surveillance* practice positions bodies under indubitable power that either/both enables and constrains bodies' thoughts, utterances, and conducts (Hardy and Thomas, 2014). It is crucial to consider that the Foucauldian conceptualisation of the panopticon goes beyond the spatial/environmental surveillance arrangements (Clegg et al., 2006) and is applicable to any technological marvel with disciplinary aims (Matthewman, 2013). Considering the context of algorithmic work regimes, one can identify specific power institutions that promise acceleration of efficiency and productivity of human employees through intelligent collaborations. These power institutions, however, simultaneously form a "net-like organisation" (Foucault, 1982), which puts employees under constant engagement with their intelligent counterparts, analyses their work interactions with algorithmic apparatuses, and engenders data clusters (Kleinberg et al., 2018), constituting the employees as *bodies of knowledge* (Heller, 1996). In other words, the novel algorithmic regimes have represented themselves as peculiar "*truths",* drawing human employees to subscribe to their computational power discourses, rendering their analytical functionality to observe the human users' interactions (O'Neil, 2016) and create disciplinary schemata to regulate the organisational actors. That being said, Foucault's notion of power is not necessarily something polemic or repressive. The exercise of power can, in fact, be a creative and productive phenomenon.

This dyadic disposition of power can be analysed from two angles within algorithmic work practices (Clegg et al., 2006). Firstly, algorithmic technologies are practices, techniques, and strategies implemented as novel organisational power discourses to bring innovative dominance modalities and/or competitive advantage over rivals (Makarius et al., 2020). Notwithstanding, the new human-algorithm/AI work partnerships are considered the most disruptive business force in the coming years (Jarrahi, 2018). Intelligent algorithms ostensibly mimic human cognition and actions and may unleash another form of creative power exercise known as "*AI Singularity"* (Upchurch, 2018). Though the prospect of facing a dominance by artificial general intelligence seems distant (John, 2017), the above

arguments regarding how intelligent agents shape, direct, and manipulate human decision-making raise caveats about how algorithmic power discourses can creatively [and covertly] dominate organisational procedures (Graßmann and Schermuly, 2021). Secondly, the relational algorithmic power discourses can alternatively become a power/knowledge discourse for the transcendence of human employees toward the Foucauldian concept of '*freedom'* (Crane et al., 2008). It goes beyond the creative dominance and objectification of human bodies through algorithmic regimes, bringing into light how algorithmic power/knowledge relations will lead to *subjectification* (Heller, 1996) of human end-users. and the extent to which subjects exercise their *resistances* (Newlands, 2021) to minimise the disciplinary measures of algorithmic power discourse (Matthewman, 2013). I argue that subscription to the Foucauldian conception of power/knowledge-subjectification corresponds to workforces' actions, reflexivity, and resistances (Townley, 1993) against algorithmic agents, giving us a theoretical edge to better understand the nuances of novel algorithmic power/knowledge discourses. In the next section, I explain the theoretical potentials of Foucault's later work on *ethics*, arguing that his particular insistence on *self-reflexivity* (Caldwell, 2007) offers a unique insight to comprehend employees' resistances against algorithmic work environments.

2.8.3 Foucault's Activism and Ethics

The third and last instalment of this theoretical lens is heavily inspired by Foucault's latest books, *The History of Sexuality: Vol 1-3* (Foucault, 2020b). In his work, Foucault ostensibly talks about his ontological project, through which human agencies constitute themselves as moral subjects (Foucault, 2020a). Foucault taps into the journey of *self-actualisation* in which subjects raise an awareness of the disciplinary procedures of power (Caldwell, 2007), seek autonomy and recognition through power dynamics (Dalgliesh, 2009), and creatively form resistances against techniques of subjection (Barratt, 2002). In a nutshell, for Foucault, an ethical subject is the one who is able to recognise and unbind him/herself from the discursive power relations (Crane et al., 2008). As discussed earlier, the human workforce is now situated in the novel algorithmic work augmentations (Allen and Choudhury, 2022) and spontaneously contributes to the cycle of power/knowledge whilst working alongside their intelligent colleagues (De Cremer and McGuire, 2022). The docile organisational stakeholders, including data scientists – or as Pachidi et al. (2021) would rather tag as "*technologists*", managers and employees – whose working lives are tangled with algorithmic tools, are all regulated subjects, constituted into the algorithmic power/knowledge apparatuses (Gardiner, 1996). However, as Foucault (2020c) argues, wherever power is exercised, there is space for resistance, and it is indeed applicable to algorithmic agents and

employees' subjectification endeavours. Knights (2002) argues that the manifestation of Foucauldian resistance in organisational work environments is beyond the conflicts and power struggle between the senior management and the employees and entails the workforce's social action to find their identity, agency, and defend their erupted dignity (Knights, 2002; Fairclough, 1993). Having said this, Foucault's problematisation of resistance has already contributed to the literature around algorithmic work commodities (e.g., Pignot, 2021; Munro, 2017; Du Plessis, 2020; Hafermalz, 2021). Yet, the main focus in the majority of those studies has been the impact of algorithmic covert surveillance – whether internally or globally – and ethics of *whistleblowing* as a creative form of Foucauldian resistance. Overall, the Foucauldian conception of *subjectification-action-resistance* (Heller, 1996) has scarcely been used in other algorithmic power/knowledge settings and/or to unfold other forms of resistance against algorithmic work regimes. For the mentioned reasons, invoking a Foucauldian lens of ethics and resistance discourse is useful to understand how employee end-users of algorithms avoid docility and domination and constitute themselves as moral bodies within the algorithmic work regimes (Crane et al., 2008).

2.9 Conclusion

To date, the literature surrounding human-algorithm partnership and AI/algorithm ethics has provided many insights around how algorithmic outputs affect the moral/ethical ethos of organisations and societies. In addition, the relevant literature offers thorough frameworks for the construction of ethical algorithmic technologies. Yet, due to the absence of empirical research from an employee/human end-user perspective, algorithm [and organisational] ethics scholarship lacks a synergistic and/or heuristic angle that entails the organisational actors' discourses of ethics and the extent to which they express and navigate their resistances against them. It is imperative to unearth how perspective, governance, and stewardship of algorithmic practices are shaped not just by those who designed them but also by those whose jobs are disrupted by algorithm transformations. As Caldwell et al. (2010) underscore, employees hold key information about the ethical implications of work practices, and it is wise to pursue, acknowledge, and value their 'say' about organisational procedures.

In retrospect, the review of algorithm literature above has revealed the discourses of power/knowledge enabled through the dawn of novel intelligent algorithms. The novel algorithmic work milieu corresponds with the Foucauldian school of thought, making critiques of how the traditional power relations are shifting from hierarchical top-down modes to more

horizontal forms, where employees are captured in the algorithmic power/knowledge discourses and rendered as docile bodies (Bergström and Knights, 2016). Human end-users are embedded in algorithmic work networks where they are contextualised, willingly self-regulated, and normalised. This is indeed the dark side of algorithmic agents that constitute individuals as ‘made subjects’ (Foucault, 1988) in a self-constructed power/knowledge cycle, in which they subtly contribute to the concept of ‘electronic panopticon’ (Lyon, 1993). Although the novel architecture of algorithmic surveillance and control takes us beyond the scope of this research, as it was explained above, there is a niche discourse where employees become aware of this digitalised (and intelligent) disciplinary mechanism, alienate themselves from it and express their resistance against it, whether individually or collectively.

In this chapter, I briefly touched upon Foucault’s analyses that underpin my research’s theoretical lens. The next chapter thoroughly examines his philosophical analyses, highlights some of the critiques Foucault received during the years and subsequently argues why Foucault’s work still promises a way forward for the exploration of algorithmic ethics. The mentioned lens and concepts in chapter 3 will later be utilised for the interpretation of findings in the discussion chapter.

CHAPTER 3: Foucauldian philosophy as a theoretical lens: A critical exploration of Foucault's work on discourse, power/knowledge, subjectification, and ethics.

3.1 Introduction

In this chapter, I provide an in-depth explanation of Foucault's philosophical work based on his visionary books and his lectures at *Collège de France* delivered in the 1970s.

Organisation and social science studies have been deeply inspired by Foucault's works since they were introduced to non-French speaking-scholarship. For example, a simple keyword search of "Foucault" in EThOS[1] (UK's e-thesis online service by the British Library) results in over 1400 hits of doctoral theses [from 1980 – now] indicating the depth, strength, impact, and applicability of Foucauldian writings that have assisted scholars in understanding the nuances of many social/organisational phenomena. Foucault's revelations, in particular, the theory of *power/knowledge* (Foucault, 1977), have made significant contributions to organisational/managerial and business in terms of theory synthesis (Moulaison et al., 2014), research design of autoethnography (Huber, 2022), and recommendations around ethical business practices (Clegg et al., 2007).

This chapter also elaborates further on the highlighted Foucauldian theories of the previous chapter in order to better structure the theoretical perspective underpinning this research. It encompasses Foucault's prominent arguments around discourse, power/knowledge, objectification-subjectification, and resistance, which are identified as the most influential Foucauldian concepts for organisational analysis (Knights, 2002). In this chapter, I highlight how Foucault's analyses will theoretically contribute to the ongoing ethical concerns around algorithmic technologies as well as signposting the most prominent criticisms around his philosophical quests. This chapter also taps into some under-represented Foucauldian notions in organisational studies, such as governmentality (Raffnsøe et al., 2019), Ethics, and Care for the Self (Crane et al., 2008), suggesting that these notions can shed further light on the theoretically overlooked notion of algorithmic ethical discourses. It is essential to mention that Foucault's later work, especially beginning from *the Archaeology of Knowledge,* shows a homogeneity throughout, making it difficult – and fruitless – to dismantle one dimension from the other. That is to say, for instance, Foucault's theory of discourse is predominantly juxtaposed with his power/knowledge theory. And subsequently, he considers power/knowledge relations, not as a 'thing' possessed or owned by individuals, but as systems of 'discourse' that could make some individuals

[1] At the time of completing this thesis, EThOS catalogue service remains unavailable due to the British Library cyber-attack in October 2023. The researcher devised this chapter before this attack took place.

disciplined and face resistance from others. Thus, in this thesis, I do not intend to select one particular dimension of Foucault's theories and loosely claim this research a *Foucauldian analysis*. But I intend to adopt a critical perspective based on his philosophical work to understand the ethical side of algorithmic work practices and, subsequently, answer the research questions of this thesis.

3.2 A brief introduction to Foucault's philosophical work

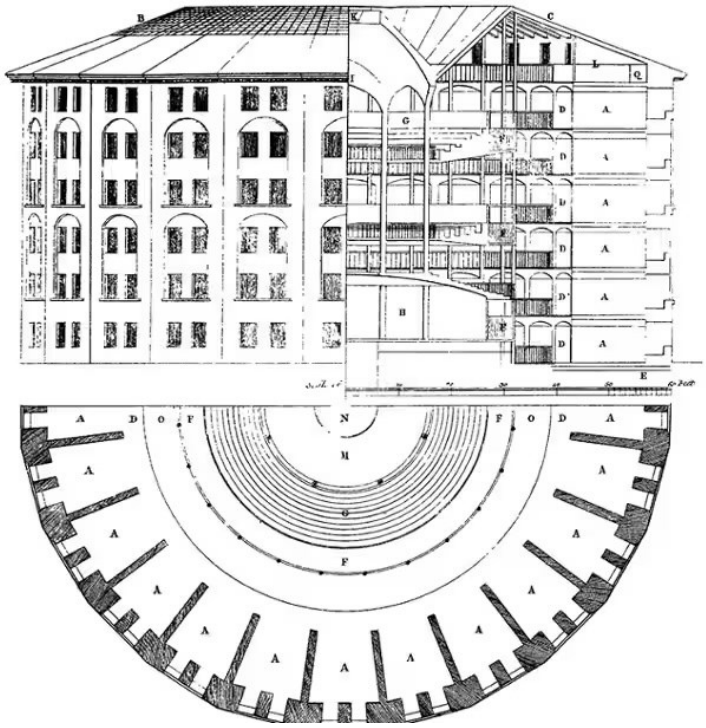
Foucault's vast corpus of writings and lectures – which were undertaken during roughly 30 years of his academic life – have been inspirational for many organisational scholars in the past few years (Townley, 1993; Munro, 2014; Knights, 2002; Raffnsøe et al., 2016; Clegg, 1994; Barratt, 2002; Hardy and Thomas, 2015; Dalgliesh, 2009; Skinner, 2013; McKinlay and Starkey, 1998).

According to Burnell (1998), Foucault's theories can be divided into two phases or analytical approaches: The archaeological and genealogical phases. The archaeological phase is mostly attributed to his earlier writings including *Madness and Civilisation* (1961), *The Order of Thins* (1966) and *The Archaeology of Knowledge* (1972). In the archaeological phase, Foucault's aim was to historically constitute the institutional and discursive systems [e.g., in examining psychoanalysis] to understand the expert's [e.g., physicians] speeches in social practices. He was pretty much interested in unpacking and analysing the governing rules and regulations within social milieux that are anonymous to social actors (Dreyfus and Rabinow, 1983). His archaeological work digs deep into reducing the subject into a function of discourses whilst treating discourse as rule-governed systems (Fairclough, 1993).

The genealogical phase of Foucault is predominantly attributed to his most influential work, *Discipline and Punish* [1977], and ultimately bolstered by the publication of *The History of Sexuality: Vol. 1-3* [1975-1986]. In this period, Foucault goes beyond the premise of structuralism and shows interest in Nietzschean genealogy (Dreyfus and Rabinow, 1983) in order to understand and articulate a link between social institutions and practices, their knowledge/truth, and theories (McKinlay and Starkey, 1998). In this phase, Foucault posits the notion of power/knowledge with a pivotal move from the concept of *sovereign power* [a type of power especially designated to the sovereign, giving them control over the life and the death of subject] to disciplinary power (Raffnsøe et al., 2019). In introducing disciplinary power, Foucault was intrigued by Jeremy Bentham's model of incarceration known as the "panopticon" [Figure 3.1] (Knights, 2002), a novel design for a penitentiary in which the inmates were subjugated/self-regulated since the 'gaze' of the central tower [the continuous existence of a prison guard] is always felt upon them (McKinlay and Starkey, 1998).

Therefore, the inmates' action and behaviour are willingly regulated, and have become docile as if they are constantly being watched, although the gaze of the supreme power might not always be present (Foucault, 1977). Although Foucault has established his power/knowledge notion based on prison as an example of an organisation, he went beyond and applied his work into other panopticon-like organisations such as barracks, hospitals, schools, asylums, etc. (Townley, 1993). He argues that the organisation and administration of human lives across various social contexts resemble prison since disciplinary power is being exercised as a practice to normalise and regulate human lives (Foucault, 1982). As he was inclined to say that power exists everywhere; not because it dominates everything, but because it stems from everywhere" (Foucault, 2020b).

Figure 3.1 Jeremy Bentham's Model of Penitentiary. Directly adopted from webpage by McMullan (2015)



It is essential to highlight two conspicuous notions within Foucault's genealogical work. First, it is not entirely accurate to label disciplinary power as a coercive, pessimistic 'thing' that is being exercised over human bodies (Heller, 1996). On the contrary, Foucault posits this form of power as productive, creative, and non-omnipotent, which is always at work and embedded in social interaction of actors (Clegg et al., 2006a). Disciplinary power, according to Foucault (1980), is not necessarily a pessimistic force aiming to oppress subjects, but rather a pervasive function that flows, resonates, and operates throughout the social body and people. In other words, disciplinary power is reinforced by the experiences,

attitudes, and knowledge(s) of the human subjects (Hargreaves, 2010). The second premise in the Foucauldian philosophy is the element of knowledge, which almost always accompanies the disciplinary type of power in his genealogical analyses. As such, Foucault conceptualises his power-knowledge as operations, relations, or functions that not only produce/devise novel knowledges around social phenomena but also translate, shape, and reconfigure the exercise of disciplinary power discourse (Hinkle, 1987). Thus, the principal in Foucault's power-knowledge is that whilst the exercise of power – in particular, the disciplinary type – regulates, controls, and normalises human lives, subsequently the people constitute bodies – or objects – of knowledge that further strengthen the power and proliferate the domain of "*Power is everywhere*" (Foucault, 1977, p. 93). However, Foucault's later projects – as I will explain in section 3.5 of this chapter – indicate that power can also be exercised by those who are self-disciplined, which indicates his theory of resistance (Heller, 1996). Overall, the concept power/knowledge is a cycle through which power has epistemic implications whilst knowledge further expands and advances the power relations (Rouse, 2005), although there is always space for human agency and resistance

With this brief introduction, the remainder of this section reviews the scholarly works surrounding the critique of Foucault's work by other critical theorists. It is with the intention to demonstrate that although Foucauldian theories have been the subject of various criticisms since their introduction, they still show scholars a way forward for the investigation of novel organisational problems. Indeed, the aim is to provide justifications that Foucault's theoretical work is still a robust analytical lens, regardless of the arguments that particular Foucauldian concepts have been over-emphasised in organisational literature (Caldwell, 2007; Crane et al., 2008).

Perhaps one of the most prevalent – or rather, conspicuous – criticisms around Foucault's work was driven by Jürgen Habermas, the German philosopher and social theorist [1929 -], a follower of the Frankfurt School and the path of Enlightenment (Fraser, 1985). The fundamental issue in their debate was to clarify whether Habermas's theory of "discourse ethics" and his strategic action of communicative reasoning offer better explanations of existing discourses of power, compared to theories such as Foucault's genealogical analysis of power (Simon, 1994). Habermas does indeed praise Foucault's successful problematisation of power in explaining its peculiar features, in particular within social institutions such as prisons or asylums, yet criticises his work as utterly non-sociological. Habermas devises and invokes the theory of communicative action, arguing that although Foucauldian power conception explains how power creatively produces knowledge and modalities, it fails to adequately demonstrate the legitimate or illegitimate use of it (Kelly, 1994). Habermas's theory of discourse ethic – structured upon communicative

reasoning – indicates that undistorted and uncoerced communications are the only true commodities for legitimate control over power and conduct (Simon, 1994). This polemic reasoning towards power has indeed made Habermas contradict Foucault's concept of power since Foucauldian power is not primarily a negative and coercive thing, or something that can be possessed, but rather a discourse to produce knowledge (Love, 1989).

Foucault was labelled as "young conservative" (Fraser, 1985) and "anti-humanist" (Caldwell, 2007) by Habermas because of his radical (and ambiguous) approaches for the development of theories to direct social change. Furthermore, Habermas followers [Foucault's critics] have argued and lamented his work's incapability in answering questions such as whether he believed in humanism. Or did he reject the idea? And if so, based on what philosophical assumptions? Yet, commentators have argued that Habermas has misunderstood Foucault's critique of modernity, subsequently accusing him of being an anti-humanist or anti-modernist (Fraser, 1985; Kelly, 1994). For Foucault, humanism accounts for values, practices, and discourses that shape the social body (Clegg, 1998), and modernity is an attitude to seek enlightenment through constant critique of our historical era (Love, 1989). For instance, in *Discipline and Punish*, Foucault is pursuing power implications in penal systems, which does underscore his quest to expand his premise of modernity through critique of historical events, rituals, and discourses. And the "critique" is indeed what the Enlightenment inaugurates in its essence (Deleuze, 2006). Hence, it is fair to deduce that Foucault's concern to bring into light the discourses, rationalities, and epistemologies of power/knowledge in penal, clinic, or other prison-like institutions indeed represents his understanding of modernity and humanism, unlike what Habermas had presumed (Knights, 2002). Foucault, of course, shifted his methodological approach – if it can be called methodology since he never provided or advocated any (Kelly, 1994) – from *Madness and Civilization* up until *Discipline and Punish*. In his analysis discussed in *Discipline and Punish* he ushers a new era in the understanding of disciplinary power and later calls for more empirical problematisation in that regard. This shift in Foucault's approach from the analysis of epistemologies towards the analysis of power caused the misunderstanding in Habermas's critique of Foucault (Simon, 1994). Habermas's critique was around Foucault's insistence on power; Foucault was wrongly accused of conceptualising power as a nonreciprocal and asymmetrical premise that exists between subject and modernity (Kelly, 1994). This argument is not entirely accurate, as Foucault never considered top-down hierarchical power. In his revelations of incarceration systems, Foucault draws attention to the notion of 'gaze', and the extent to which subjects make objects of knowledge and perpetuate the continuity of particular power relations and institutions (Barratt, 2002). For instance, Foucault points out how human bodies (e.g., delinquents, madmen, soldiers, and

pupils) actively and willingly participate in the processes which establish power organisations such as prisons, military barracks, asylums, and schools, as well as novel sciences of penology, military, psychoanalysis, and education (Foucault, 1977). In other words, subjects give meaning to power exercise since they are constituted as objects of knowledge/truth concurrently (McKinlay and Starkey, 1998).

With this in mind, linking Foucault's perspectives of modernity and techniques of rationality to the emergence of novel 4.0 technological breakthroughs in organisations (Weiskopf and Hansen, 2023) is the core focus of this research. Indeed, the conceptualisations of *Critical Theory* from the Frankfurt School (Wiggershaus, 1994) – mostly attributed to Adorno and Horkheimer, and its defender and critique, Habermas (Hohendahl, 1985) – might be insightful for the theoretical investigation of algorithmic work, especially from an emancipatory or subjects' liberty lens. Yet, the Frankfurt School is predominantly structured on Marxian criticism of capitalism and states many pessimistic [or rather negative] deductions around the transcendence of human subjects (Kelly, 1994). Said differently, the Frankfurt School's offers a controversial view for 'subjects liberty or emancipation, deeming it to be impossible without social totality or the dichotomy between subject and object (Hohendahl, 1985). Although Habermas's attempt to reconstruct the older generation of the Frankfurt School by integrating linguistic, social action, and system theories shed light on subjects' resistances against the totality of administered societies (Hohendahl, 1985), compared to Foucault's theories, it categorically fails to manifest the power relations that produce discourse of truth or knowledge (Ahonen et al., 2014).

Having said this, the literature surrounding the adoption and implementation of algorithm technologies abandons the dominant discourse of power (Neyland and Möllers, 2016) and highlights a theoretical void regarding the subject's perspective of the algorithm's ethical dimension. Although the existing literature has advanced theoretical backgrounds in relation to algorithmic agnatical power (Peeters, 2020) or the algorithmic surveillance power (Newlands, 2021), it has largely neglected the existence of power discourses amongst the key organisational subjects (Moran and Shaikh, 2022), who either influence the adoption decision, or are the human end-users of algorithmic agents (Bader and Kaiser, 2019). And this is where Foucauldian concepts show a supremacy compared to other [post]structuralist philosophies, since not only does his work look into how discourses shape and transform the context of new organisations, but it also incorporates human agency, subjectivitym and resistance (Barratt, 2008); relevant – yet overlooked – issues in the algorithmic work literature (Pignot, 2021; Cameron and Rahman, 2022). Therefore, adopting a Foucauldian perspective will not only help the thesis advance the existing literature on AI/algorithm ethics,

exploring nuances such as ‘power’, but also provide an analytical lens to uncover the concepts of human agency and resistance in algorithmic work practices.

The remainder of this chapter is structured as follows. First, I explore and elaborate on Foucault’s theory of critical discourse and its relevance to the context of algorithmic work processes. Second, I tap into his genealogical work around power/knowledge, with the aim of justifying that his conception of power/knowledge is not only able to explain the power relations and interactions between humans and algorithms but also to uncover disciplinary power discourses and dispositives (Raffnsøe et al., 2016) that drive the adoption and implementation of algorithm technologies. Third, I will discuss the later Foucauldian inheritance, in particular his project related to ethics, care for the self, which alludes to his quest in the reconfiguration and reinvention of current organisational orthodoxies. The aim is to constitute a theoretical perspective that is capable of unpacking the ethical dimensions of algorithmic tools, with the emphasis on finding the discourses that flow amongst subjects.

3.3 Foucault’s Theory of [Critical] Discourse

The method of discourse analysis is very much influenced and guided by structuralist theory thinkers, in particular, Louis Althusser (Althusser, 1969), Jacques Derrida (Derrida, 1978), as well as their follower, Michel Foucault (Foucault, 1972). Amongst these, Foucault’s interpretations have emerged as an insightful analytical approach to investigate ‘truth(s)’ and politics within language (Graham, 2011). Foucauldian theory of discourse analysis is not concerned with the intangible meanings of utterances or discursive statements, but it aims to understand the functionalities and systems within language that shape the ‘social body’ (Hardy and Thomas, 2014). The premise of discourse appears in different aspects of Foucault’s work; yet it is unclear and cryptic (Burrell, 1998). For Foucault, discourses are not simply utterances that describe objects or the world; they are able to constitute them by materialising phenomena through the approaches in which they can be categorised and made sense of (Hardy and Thomas, 2015). In other words, Foucault sees statements, enunciations, and formations not for what they say but for what they are able to *do* (Bergström and Knights, 2006), what ethical or political effects they have, and what function they serve (Graham, 2011). As it was mentioned in the previous chapter,

Foucault, throughout his life, was not vexed by any specific type of organisation or power institutions, yet his concern was around the organisation or administration of lives (Mennicken and Miller, 2014). His concern also includes the constitution and configuration of individuals, not through the micro-semantics of language, but through the functionality of discourses that *make up* subjects (Fairclough, 2003). Hardy and Thomas (2015) explain

further the Foucauldian conceptualisation of discourse in order to signpost his core concern of materiality within discourses. They argue that although Foucauldian discourses explore the formations and functions of text and language, they also acknowledge the ways in which discourses structure material principles, institutions, and overall, the material world (Hardy and Thomas, 2015). And in juxtaposing discourses with the material world, Foucault ultimately brings to light the notion of power relations, arguing that the exercise of power is the true embodiment of materiality, physicality, and corporeality (Foucault, 1980). Indeed, a dyadic understanding of textual meanings and material exercise of power in which those Foucauldian meanings are entwined will help the scholarship to better explore and critique the dynamics of power relations in organisational contexts (Hardy and Thomas, 2015).

In the context of this research, Foucauldian theory of critical discourse would be beneficial in at least two ways. Firstly, by integrating his theory of critical discourse, this research will be able to understand the dominant power relations through discourses that influence and characterise (Hardy and Thomas, 2014) the adoption of algorithmic tools in work practices. As it was argued in the previous chapter, the introduction of algorithmic technologies is aligned with the promise of automation and augmentation (Eglash et al., 2020; Grønsund and Aanestad, 2020) through the decision-making discourses that are very much directed by top executives (Spencer et al., 2012) and/or technological elites (Dwivedi et al., 2021). According to Foucault (1980), meaningful practices and strategies are built within discourses to make the objects and subjects 'known'. As such, the entire process of algorithm design, adoption, and implementation can be explored through the Foucauldian theory of discourse with a view to distinguish the relations, impacts, and complexities of power and how such discourses shape knowledge. Indeed, as many scholars have already argued, Foucault never saw power as a "thing" that can be possessed by individuals, but rather as a transformative discourse able to produce things (Heller, 1996; Crane et al., 2008). Thus, in the context of this research, I will not ask *who* has the power to drive decision-making for algorithm adoption or what *typology* of power is in place, as it has been the case in the prevalent power categorisation theories such as the one by French and Raven (Elias, 2008). Instead, I will be looking into how the relevant discourses harbour or translate power relations in algorithm adoption processes; how the key stakeholders' perspectives, languages, and conducts are enabled or constrained (Hardy and Thomas, 2014) by the discourses of power in the context of the algorithmic work environment. And how the networks of power discourses exercised by individuals shape and/or make impacts on the algorithmic work milieux.

Secondly, going further from the identification of discourses around power, Foucauldian theory of discourse assists this research in the exploration of ethical discourses

in algorithmic work. As Burrell (1998, p. 16) mentions, Foucault's prevalent work on discourse – *The Archaeology of Knowledge* – indicates that truths are generated through various systems of discursive formations [i.e., statements] and discourses and are independent of the speaker's consciousness. It means that considering discourses as autonomous discursive statements limits our understanding of the diversity and heterogeneity of discourses (Raffnsøe et al., 2019); rather than accepting one single truth that can be accepted as the representation of reality, Foucault encourages us to look deeper into how multiple discourses produce, reject, and influence social structures and interactions (Bergström and Knights, 2006). The ethical issues around AI/algorithm technologies and recommended ethical frameworks explained in the literature highlight the intensification of power discourses (Hardy and Thomas, 2014) driven by ethicist scholarship (Tsamados et al., 2022). The majority of ethical discursive formations [dominant discourses] in the relevant literature particularly show a single truth taken to be accurate since they are based on the suggestions of data/IT elites. This is indeed in contrast with the Foucauldian philosophy of discourse that rejects a single universal truth and embraces multiple truths (Fairclough, 1993). Of course, the ideas that lead to the adoption of algorithms generally stem from tech elites' strategies (Jarrahi, 2018), who are also praised for their awareness of algorithmic ethical challenges. Yet, the partnership of human employees with the algorithm tools (De Cremer and McGuire, 2022) is a novel power modality in which all individuals actively participate in power discourses, constitute bodies of knowledge, and produce discourse of ethics (Weiskopf and Hansen, 2022) that may differ from those of data elites. A key aspect that is overlooked theoretically within the algorithm ethics literature. Furthermore, the prominent existing theoretical frameworks [e.g., TOE, TAM] that tap into discourses and decision-making processes of algorithm adoption, though profoundly paving the way for effective introduction of algorithms, theoretically demonstrate weakness around ethical discourse of algorithms.

3.4 Foucault's Conception of Power/knowledge

The premise of power is defined from two major perspectives within organisational literature. The first view emerged out of structuralist-Marxian philosophy that sees power as a negative coercive force possessed by a group of elites (Clegg et al., 2006b). Following the structuralist foresights, Max Weber saw power in relation to economy, control, and the dynamics of production (Heiskala, 2001). As he conceptualised, power relations emerge not only from the individuals who have the ownership of production but also those who have the knowledge and competencies on how to operate, control, and/or govern the means of production [in contrast with Marx's capitalistic view of power]. Weberian thought rather puts

power as “an ability to get others to do what you want them to, against their will if necessary (Weber, 1978). This view of power – although enacted as a starting point in studying power – has been challenged, critiqued, and reconfigured by many scholars over past decades (Campbell, 2009; Heiskala, 2001; Jimenez-Anca, 2013; Uphoff, 1989; Clegg, 1994; Raffnsøe et al., 2019). However, a great proportion of knowledge around the problematisation of power in the 20th century – particularly in relation to organisational life – focuses only on illegitimacy and dominance by power (Pfeffer, 1992) or the extent to which exercise of power acts against democracy in favour of dominance (Perrow, 1979). Yet such narrow focuses from organisation theorists only scratch the surface of micro-strategies of power and arbitrarily consider power as something that aims to impose subordination and docility (Clegg et al., 2006a).

With the emergence of postmodernism and post-structuralism [though he mentioned myriad times that he never considered himself as a poststructuralist/postmodernist (Heller, 1996)], and subsequently the way Foucault challenged the existing theories of power, the understanding of power started to shift significantly. As such, Foucault examines the techniques of power through their impacts, not through power’s existentiality. In other words, the focal point in his analysis is to understand ‘how’ power relations work and ‘how’ subjects react to the exercise of power.

Perhaps one of the paramount aspects in Foucault’s conception of power is where he shifted attentions from sovereign power orthodoxy towards the premise of disciplinary power and consequently biopower/biopolitics (Caldwell, 2007). Foucault envisages sovereign power exercise as a feature of monarchical regimes, or rather sovereigns, which manifests itself through the ritual of torture or public executions, in order to impose corrective measures on subjects’ behaviour (Foucault, 1977). Indeed, for Foucault, sovereignty and its power relations are repressive, prohibitive, and ultimately negative, with the aim to impose law and regulation at any cost (Lilja and Vinthagen, 2014). As the paradoxical dimension to sovereign power, Foucault introduced ‘disciplinary power’ in his genealogical *Discipline and Punish,* which he signifies as a productive or creative institution (Foucault, 1977). In order to constitute his hypothesis of disciplinary power, he moves away from the Marxian or Weberian questions on what power is or who has the power and considers power as active and inescapable (Clegg et al., 2006a) and asks what the role, intention, and impact of power relations are. Furthermore, along with the premise of disciplinary power, Foucault has introduced two key elements in order to interpret his historical shift from sovereign to disciplinary power: panopticon and governmentality. I will further explain these two notions as they have ramifications around the emergence of algorithmic technologies for organisational processes and their ethical discoursers.

As it was argued earlier in this chapter, Foucault was intrigued by Bentham's design of penitentiary system known as the "Panopticon" (Clegg et al., 2006b). The key idea of the panopticon is surveillance through constant inspections of inmates in a way that the subjects are not necessarily able to see their incarcerator (Foucault, 1977). Bentham consequently expanded his design to other forms of social institutions such as hospitals, factories, and barracks, where there is always one superintendent responsible for overseeing human bodies or people (McKinlay and Starkey, 1998). Drawing on Bentham's view, Foucault conceptualised the idea that panoptic prisons resemble organisations in the era of modernity in which subjects are dominated, self-disciplined, and self-regulated (Barratt, 2002). More importantly, Foucault integrated Bentham's panopticon in his power/knowledge revelations, arguing that power no longer exists in relation to the place, rules, or even in the possession of individuals, but rather is a decentralised feature through surveillance of individuals to normalise their conduct, attitudes, and traits (Townley, 1993). Foucault ultimately suggested that the power exercised via surveillance and control over human bodies and souls not only would constitute docile social actors but also make each human subject an object of knowledge (Foucault, 2020a). The human objects of knowledge will subsequently contribute and proliferate the existing knowledges around how panoptic practices can advance in modern organisations (Heller, 1996).

Indeed, the existing literature has touched upon the panoptic feature of algorithmic tools and the extent to which the algorithmic gaze/surveillance targets human players not just inside organisations (e.g., Newlands, 2021; Pignot, 2021; Walker et al., 2021) but also outside, encompassing global surveillance of citizens (e.g., Neyland, 2015; Munro, 2018). However, it seems the existing literature has confidently subscribed to polemic side of the algorithmic gaze and is content with the discoveries around illegitimate surveillance regimes or misuses of algorithmic power. The literature on algorithmic surveillance has overlooked the human side within algorithmic panoptic control that goes beyond the tangible dimensions of algorithms. It has also neglected panoptic procedures that circumvent subjects within algorithmic work practices. Such panoptic power/knowledge procedures are not necessarily steered by algorithmic surveillance but by managerial standardisation and normalising discourses that promise consistency and prosperity via algorithmic tools (Allen, 2019). Invoking Foucauldian concepts of surveillance and control, Barratt (2002) argues that the exercise of power in modern organisations is no longer through physical observations and control over bodies/souls, but rather through disciplinary [and incognito] micro-techniques such as employee upskilling programmes, performance evaluations, and ultimately, *managerialism* (Grey, 1996). Indeed, novel people management methods, career development schemes, and fundamental decision-making processes in organisations are all

reminiscences of Foucauldian panoptic control and dominance over bodies (McKinlay and Starkey, 1998). In the context of algorithm work practices, employee end-users are embedded in an algorithmic panopticon where the decision-making practices have already advocated the adoption of algorithm agents. Not particularly because decision-makers are in the position of power, but because the new dawn of modernity is aligned with the introduction of novel technologies that perpetuate disciplinary power/knowledge mechanisms (Weiskopf and Hansen, 2023). And employees are asked to interact with these tools, and those interactions are meticulously recorded, monitored, and analysed to further the position of the algorithmic power/knowledge cycle. In essence, the introduction of algorithms and human end-users' interactions with them highlights a distinctive administration of lives, or "governmentality" (Introna, 2016), which will be explained in the following.

The second Foucauldian premise, which needs introduction and has importance in the context of algorithmic work processes, is *Governmentality* (Foucault, 1980). He used this term initially in his lecture at *Collège de France* as a product of neoliberalism and to justify that government is a form of power that entails conduct upon the conduct of others or simply put, to govern well by governing less (Mennicken and Miller, 2014). Foucault explains that domestic or global issues (e.g., unemployment, economic recessions, or even pandemics) are pillars that shape and represent governments. Specific governmental rationalities are in place to diagnose or problematise the social issues and achieve their objectives (Weiskopf and Hansen, 2023). The government rationalities are also aimed at regulating, normalising, and subjugating bodies and souls of subjects (Flyvbjerg, 1998). Yet Foucault's concern was not just to explore new governance discourses or bureaucratic administration of lives, but also the impacts of those discourses and how they enable subjects to act freely whilst still being subjugated and normalised (Leclercq-Vandelannoitte, 2011). In other words, Foucauldian governmentality is not solely about administrative strategies of subjects in a broader sense; but it includes how those subjects, *in their own autonomy,* act upon the administrative strategies to seek or shape their lives (Oakes et al., 1998). Foucault's insistence on a subject's liberty and autonomy in rationalities of governmentality indeed follows his earlier conception that sees power-knowledge not as exploitative or prohibitive mechanisms, but rather as a creative and facilitative phenomenon (Clegg et al., 2006a). In line with Foucault's concept of creative governmentality, Weiskopf and Hansen (2023) draw attention to the digital data-enabled governance technologies and practices that render human subjects as objects of knowledge whilst being normalised within their work/life assemblages. The premise of *algorithmic governmentality* (Cooper, 2020) thus incorporates governance rationalities and practices that are structured upon massive data clusters, with the aim to produce disciplined subjects. Algorithmic governmentality accounts as a novel

form of disciplinary power-knowledge to extrapolate citizens' behaviours [human bodies as subjects of knowledge/truth] and reduces them to subjugated individuals (Weiskopf and Hansen, 2023). Particularly, algorithmic governmentality opts to regulate human subjects' conduct in the sense that they have liberty to choose between appropriate or inappropriate behaviours (Raffnsøe et al., 2019). Although the existence of human subjects within algorithmic governmentality differs from earlier liberal governments, where they are considered cogs or gears in the machinery of production (McKinlay and Starkey, 1998), it is still a reminiscence of subjectivity enslavement and volunteer obedience to governmental rationalities (Weiskopf and Hansen, 2023).

It is crucial to emphasise that the notion of Foucauldian governmentality [also the notion of algorithmic government] and the concept of panopticon surveillance have been predominantly utilised only to explore the power/knowledge discourses that create subjugated individuals in various organisational or societal contexts (see Barry, 2019; Introna, 2016; Isin and Ruppert, 2020; Roberts, 2019; Galière, 2020). However, the space and discourses of ethics seem to be theoretically underexplored in algorithmic work practices [or rather governmentality], specifically through Foucauldian philosophy (Weiskopf and Hansen, 2023). Moreover, the studies focusing on the subjects' marginalisation within algorithmic governmentality offer a holistic view of the concerning individuals both inside and outside organisational contexts. This is clearly providing an obscure picture of how ethical discourses are traversed via organisational actors (Crane et al., 2008), and how the ethical subjects are forged/constituted (Cooper, 2020) across algorithmic organisational milieux. This theoretical gap, of course, overlaps with Foucauldian analyses of objectification, subjectification, and resistance, which are explained in detail in the following section.

3.5 Objectification, subjectification and Foucault's Theory of Resistance

Foucault argues in detail how power and knowledge produce and reconfigure each other. In other words, no power mechanism exists without the circulation of fields of knowledge, and no knowledge would be constituted if there existed no power relations (Foucault, 1977). As it was argued, the shift from sovereign power to disciplinary power was a pivotal point in Foucauldian genealogical analysis (McKinlay and Starkey, 1998) where the discourses of power/knowledge were fundamentally transformed from a force of restrictions to a form of creative self-regulation (Clegg, 1994). With this in mind, Foucault explains this power/knowledge evolution through concepts of *objectification* and *subjectification,* and upon the discourses that affect subjects/people in their contextual life. Yet before showcasing how this Foucauldian shift is applicable to the topic of algorithmic work, it is worth clarifying some

imperative terms being used in Foucauldian social constructionist philosophy. Interestingly, the body of literature surrounding business and management has hardly integrated the conceptual shift from objectification to subjectification. And the majority of studies that have been keen to bring together those concepts into their research area are situated in humanities discipline, more specifically feminist theory (see, for example, Staunæs, 2003; Gressgård, 2013; Ussher and Perz, 2020; MacDonald et al., 2006; Gill, 2007). Thus, I will rely on the feminist researchers' interpretations to explain how Foucault invokes those in his philosophical revelation of disciplinary power/knowledge and aims to link them to algorithmic organisational processes.

The polarities of objectification versus subjectification are often aligned with how an individual is able to understand and acknowledge their value (Schraub, 2016). An objectified person – particularly in feminist theories – is the one who is being controlled, instrumentalised, and accepts quantifiable facts and knowledge(s) (Hart and Fuoli, 2020; Nissen, 2003). Expressed simply, an objectified person is valued because of their functionality, and can always be interchanged with other objects of the same type (Gill, 2007). Nussbaum (1995) illustrates the instrumentality of an objectified person with an analogy of a pen or a word processor; all of them function to fulfil the need of transcribing words on a page and can be replaced whenever necessary. Subjectification, on the other hand, stems from the notion of subjectivity. For Staunæs (2003), subjectivity stands as a post-structuralist premise for an individual's sense of self, which encompasses stability along with transformation and ruptures of the character. As such, in subjectification, the human subject seeks dignity, recognition, and value through the eyes of others (Ruppert, 2008). Reiterated simply, the subjectified people are neither denied nor enslaved but rather are not acknowledged and/or valued as instruments in relation to other members of the society. In Foucauldian philosophy, however, such conceptions are embedded and explored in relation to power/knowledge discourses (Clegg et al., 2006a). The concept of subjectivity in the Foucauldian project, though overlapping with feminist theories, describes individuals who are either subjected to or targeted by power via dynamics of power/knowledge (Dreyfus and Rabinow, 1983; Foucault, 1980). The power/knowledge discourses then transform individuals into subjects who possess a sense of meaning, aim, and reality while actively participating in the discourses of power/knowledge. For Foucault, individuals are both objectified and subjectified by power exercises and simultaneously undergo and exercise power and produce knowledge (Foucault, 1980). Heller (1996) offers a more detailed explanation on the process of subjectification through Foucauldian power/knowledge theory. He argues that Foucault's process of subjectification is a heterogeneous one that is more about *how* subjects are positioned within counter-hegemonic discourse and *how* this

positioning or existence constitutes alternative or rather new discourses and social formations. Thus, Foucauldian processes of subjectification is not merely about whether subjects are constituted by mechanisms of power/knowledge (Heller, 1996), but how those subjects assume their agency or authorship in the constitution of a discourse or an ideology (Knights, 2002). As Bergström and Knights (2006) suggest, Foucauldian subjectification is all about convincing individuals – those objects of power/knowledge – that they are willingly choosing a change, without actually imposing a change!

In light of the processes of adoption and implementation of algorithmic technologies, the data engineers and IT elites offer a new discourse as '*elevating organisational productivity and performance as well as cost savings*' through the utilisation of algorithmic technologies (Sestino and De Mauro, 2022; Pillai and Sivathanu, 2020). At the outset, the emergence of algorithmic work regimes [discourses] has been steered not solely by data scientists, AI ethicists, and technologists (Pachidi et al., 2021) but also approved and praised by top management (Dwivedi et al., 2021) and disproportionately welcomed by human employees (De Cremer and McGuire, 2022). Algorithmic technologies are introduced with the promise of swift data processing and enhanced prediction [autonomous decision-making] (Jarrahi et al., 2021) gradually abate traditional work discourses whilst human employees are constituted as objects of study to strengthen the algorithmic prediction features (Vassilopoulou et al., 2022). This is in line with Foucault's (1980) revelations around objectified individuals and the processes of subjectification. Organisational actors who are situated in the algorithmic power/knowledge discourses can be categorised into three groups. The first are those individuals who are constructed by algorithmic practices and are objectified by the existing power relations, such as employee end-users who are subjugated by algorithmic practices. The second group is those in the process of subjectification, which incorporates many stakeholders, including data scientists and engineers, top executives, and/or even employee end-users. These particular organisational players are those who predominantly make adoption decisions, offer technological, managerial, or employee support, and have willingly accepted algorithms as their work partners; those whom Foucault postulates as the subjects that actively situate themselves in the power/knowledge discourses and structuralise novel subject positions or discourses (Bergström and Knights, 2006).

Naturally, the subjectified individuals would further perpetuate the novel disciplinary mechanism of algorithmic work since they inevitably and reciprocally have become the components in the socially constructed algorithmic work (Fairclough, 1993). The third group, however, are those who are able to eliminate and/or modify the disciplinary mechanism of power relations and express their resistance against the algorithmic work regimes (Weiskopf

and Hansen, 2023). The current literature has indeed highlighted the issue of workforce resistance against algorithmic technologies, yet the debate is predominately shifted towards algorithmic surveillance of gig economy business platforms (Newlands, 2021; Pignot, 2021), overlooking the nuances of resistance that links to power relation discourses in adoption of AI/algorithms (Burton et al., 2020) or subjects' expressing concerns around algorithmic ethical issues (Li et al., 2022). Human workforce resistance against algorithmic indeed overlaps with Foucault's conception of resistance [which resonates around power/knowledge] and can be explored through his analytical lens. In the following, I elaborate on the notion of resistance through Foucault's power/knowledge project, which has been a pivotal issue for many organisation scholars (Knights, 2002; Mennicken and Miller, 2014; Barratt, 2002; Raffnsøe et al., 2019; Clegg et al., 2006b; Heller, 1996) and seems timely and pertinent to algorithmic work contexts (De Vaujany et al., 2021).

For Foucault, the chief concern was to unpack how power/knowledge relations are at work to discipline the population and subjects and how epistemological rules and rituals [or *savoir*-knowledge] are formed through subject disciplining (Knights, 2002). However, Foucault argues that wherever there is power, there is also space for resistance (Clegg, 1989). In other words, subjugated individuals strive to free themselves of the disciplinary panopticon-like practices and seek autonomy and agency (Barratt, 2008). Resistance is conceptualised (Foucault, 2020b) as an opposition against, or rejection of, the institutional power discourses that constituted them as subjugated, docile, and normalised bodies. Foucault's revelations of resistance were predominantly drawn from his later work, specifically around control over subjects' sexuality and the extent to which individuals oppose/reject the stigmatisations, labelling, and discriminations against them (Foucault, 2020b). However, the Foucauldian concept of resistance has been criticised as a fruitless effort with limited/poor outcomes for human subjects' freedom (Caldwell, 2007). In here, I invoked organisational scholars in order to justify that not only is Foucauldian resistance not a dystopian premise and offers the possibility for subjects' liberty and emancipation (Heller, 1996), but also it provides theoretical insights for the current human-algorithm work interactions and understanding the ethical side of algorithmic work.

Looking into postmodern organisational envisions of resistance, Knights (2002) argues that although the embodiment of resistance is linked to the subject's social identity, rather than the pursuit of dignity and autonomy, the latter concerns are not entirely irrelevant, particularly in the midst of facing a threat to the erosion of 'self'. Organisational discursive and normative practices are intrinsically aimed at regulating their employees (Mennicken and Miller, 2014). Subsequently, the dominated subjects would reject or oppose the systems of power once their autonomy and dignity are in the condition of decay (Clegg, 1998). Clegg et

al. (2006b) and McKinlay and Starkey (1998) argue that although it seems the goal of resistance is to overcome the normative control and prevail over the power/knowledge, resistance hardly ever leads to such power transformations. That is to say, contrarily to what is expected, employees' resistance ubiquitously reinforces power dynamics (Barratt, 2002). To unpack this surprising phenomenon, first we need to acknowledge that on the one hand, the employee subjectivity in the workplace emerges from the managerial power/knowledge programmes (Clegg et al., 2006a). On the other hand, the reinforcement of power emanates from the employee's identity and sense of belonging. Therefore, they are unable to alienate/detach themselves from the disciplinary power mechanisms irrespective of resistance discourses. Put differently, the managerial disciplinary orthodoxies [the power/knowledge discourses] are enacted with the utmost totality that makes employee resistance [activism] constrained (Dreyfus and Rabinow, 1983). For this very reason and the polemics of power from Foucault, scholars have wondered why *resist* when it seems such a futile effort (Clegg et al., 2006a). Indeed, Foucault's demise and his unfinished ethical/aesthetical project might have later clarified matters for Habermas, impelling him to label Foucault as *anti-humanist* (Knights, 2002). In order to clarify this misconception, we need to look into the *reversibility* of the power/knowledge grid (Foucault, 2019). In that regard, Heller (1996) argues that the power/knowledge mechanisms are indeed shaped and institutionalised by discourses from both groups of dominant and dominated individuals, thus cannot be divided into subgroups of included and excluded discourses. Furthermore, the process of subjectification does not necessarily restrict liberty via power. Subjectification enables those subjects within the gird of power/knowledge to speak and voice their opposing discourses and/or choose other tactics to minimise the dominating intentionality of power relations (Foucault, 1990).

Stating this, we can now go back to the earlier argument that challenged Foucault's resistance as a poor prospect and rely on his own interpretations [found in his under-represented lecture] to illustrate why scholars were wrong. For Foucault, power is neither evil nor destructive nor always dominative (Bernauer and Rasmussen, 1988). And of course, subjects can embark on an opposite subjectification journey [instead of being the subjects of power relations (Bergström and Knights, 2006)] and resist power relations through reversibility or flexibility of power relations. Foucault has used two instances to explain this difference: the exercise of power through love/pleasure as well as pedagogy and the transmission of knowledge (Heller, 1996). In these two examples, Foucault sees no domination of subjects but liberation in the sense that power is equipping people with the capacity to reconfigure the conduct of others, and it is being used to proliferate liberty (Foucault, 1982). Ultimately, the utopian dimension of Foucauldian power/knowledge is not

about how fruitful, fruitless or influential the resistance is in shifting power exercises but is rather about acknowledging that there is always a *possibility* and *space* for freedom (Dalgliesh, 2009) depending on *how* power is being exercised (Heller, 1996).

With the above justification, I now use and elaborate on how Foucault's space for resistance links to novel power/knowledge discourses of algorithmic work practices and how it can provide theoretical insights on the topic of its ethics. As it was discussed in the previous chapter, perhaps the most pivotal reason for the adoption and integration of algorithms is to augment or automate work practices that are traditionally undertaken by human employees (Jöhnk et al., 2020). That is to say, algorithmic power/knowledge relations are in place to maintain productivity and work consistency (Campbell et al., 2020). As the literature review illustrated, work placements and partnerships of human employees with the algorithmic technologies, although proven to flourish the organisational efficiency (Sutton et al., 2018), have faced resistance by the employees at different levels of hierarchy (Velkova and Kaun, 2021). Yet, the existing studies lack theoretical, empirical, and contextual exploration around the issue of resistance to algorithmic power/knowledge work practices (Bucher et al., 2021; Newlands, 2021), let alone fathom how and why any resistance would take place. Utilising Foucault's analyses, however, would be beneficial in several ways. Firstly, as it was discussed earlier, Foucault sees power not as a coercive or destructive force but rather as a mechanism of intentionality and creativity (Heller, 1996). Through his lens, not only does this research theoretically investigate the reversibility/flexibility of algorithmic power/knowledge discourses but also the space and possibilities of resistance. Indeed, Foucauldian analyses would also help to dig deeper into power relations that influence the introduction of algorithm tools and the production of their ethical discourses. His critical work would also help to grasp whether the exercise of power is a utopian, liberative one that supports employee liberty and emancipation or a dystopian one with the aim to dominate human subjects through its computational supremacy.

Foucault's theories give this research an edge – particularly with regards to theoretical backbone – over similar studies as it also examines the discourses of power/knowledge that affect the ethical dimensions of algorithmic tools. As a review of the relevant literature indicated, the knowledge around AI/algorithm ethics predominantly elaborates on the technological side of the machine (e.g., Tsamados et al., 2022; Mittelstadt et al., 2016; Amitai and Oren, 2017; Russell et al., 2015; Yu et al., 2018) whilst overlooking the discourses of employee subjects who may constitute their own conceptions of ethics through subjectification, activism, and perhaps even their resistance (Caldwell, 2007) against algorithm work regimes. Having said this, Foucault also talks immensely around the truth of

ethics and the evolution of ethical subjects in his later work, of which I will discuss more in the final section of this chapter.

3.6 Foucault's work on Ethics and Care for the Self

Foucault's concern for human subjectivity and resistance against power/knowledge apparatuses finally led him to analyse the conditions and possibilities for human 'liberty' (Knights, 2002). Foucault begins his work on the ethics and aesthetics of the subject through the analysis of 'sexuality' and genealogical comparisons of Roman-Greek culture with pagan and orthodoxies of Christianity (Gardiner, 1996). Through his analysis, he explains that the dawn of western modernity is attributed to imposing great normative power relations on people's sexuality (Foucault, 2019). And as such, people's sexual conduct was the topic of observation [disciplinary 'gaze'], categorisation, and evaluation, and consequentially being the object of knowledge: any problematic, irregular sexual practice was subjected to control and correction (Foucault, 2020a). Eventually, Foucault makes a controversial – yet sound – differentiation between antiquity and modern societies in terms of how subjects' sexuality was treated: In the age of antiquity, as Foucault elaborates, a moral and virtuous individual was recognised by his/her own search for 'ethics of exitance, or simply saying, his/her ability to seek liberty and self-recognition, pertaining to their sexual behaviour (Barratt, 2008). Thus, in the ancient Greek/Roman societies, the concern for morality or ethics was not around what is deemed as appropriate and/or prohibited conduct, nor was it about how to be obedient subjects. But rather around how subjects are able to add to their own lives, change or transform themselves, and be able to stand against the rules/codes that determine morality and ethicality (Crane et al., 2008). In other words, an ethical subject for Foucault is the one who is the potential source of liberty, spontaneity, and selfhood, and not the one who is constituted by power/knowledge mechanisms (Gardiner, 1996).

The ethics and aesthetic subject dimension of Foucault's project, unfortunately, remains incomplete, and it is unclear how to allude to his historical analogy to the meanings of autonomy and liberty of the present. His death indeed made it difficult to speak of his true allusions (Barratt, 2008). Yet one can still detect elements of agency, self-reflexivity, and actualisation in his later recommendations (Townley, 1995); highly relevant notions to contemporary social formations, including organisational milieux (Caldwell, 2007; Clegg, 1994). And for this very reason, many organisational scholars have encouraged further research to explore novel business, management, and economy issues of organisations via Foucault's conceptualisation of ethics (see e.g., Caldwell, 2007; Crane et al., 2008; De Vaujany et al., 2021; Ganti, 2014; Munro, 2014; Raffnsøe et al., 2019; Weiskopf and

Willmott, 2013; Barratt, 2008; Alakavuklar and Alamgir, 2018). The ethical dimension of Foucault's philosophical thought is particularly aligned with notions such as acts of parrhesia [fearless speech or expression of truth], aesthetics, and care/technologies of the self (Raffnsøe et al., 2019). Undoubtedly, such manifestations in Foucault's thought allude to the situations where subjects become the authors of their own lives in the contemporary power/knowledge relations (Barratt, 2008).

In order to illustrate how subjects become autonomous in the power/knowledge cycles, Foucault invokes a complicated process through which subjects exercise control over themselves. This type of subjectivity, as Foucault calls "*modality of relation to self"* (Foucault, 2020a), entails the constitution of human individuals who are social and juridical subjects. The subjects who respond to disciplinary power relations and do not escape from them are those who actively and willingly participate in power relations because power has given them freedom (Raffnsøe et al., 2019). Thus, *technology of the Self* or *Modality of relation to Self* (Foucault, 1988), does not mean that subjects are to be dominated by normalising rules or codes of conduct, but how they can devise their own version of codes of conduct (Crane et al., 2008). In Foucault's (2019) own words, an ethical individual is shaped through:

"A process in which the individual delimits that part of himself that will form the objects of his moral practice, defines his position relative to the precept he follows, and decide on a certain mode being that will serve as his moral goal. And this requires him to act upon himself, to monitor, test, improve, and transform himself. A moral action tends towards its own accomplishment; but is also aims beyond the latter, to the establishing of a moral conduct that commits an individual, not only to other actions always in conformity with values and rules, but to a certain mode of being, a mode of being characteristic of the ethical subjects [p. 38]*."*

Although criticisms have been raised in the sense that Foucault neglected the concepts of unity, dialogue, and collectivism in the development of ethical subjects (Gardiner, 1996), Crane et al. (2008) argue that Foucauldian ethics is indeed informative for self-improvement and for the creation of a better society. On this occasion, Foucault's ethical framework incorporates the capacity to transfer the knowledge/truths on how the analysis, critique, and actions can be implemented to minimise the domination individually and collectively and by organisational actors (Raffnsøe et al., 2019). This type of approach clearly opens up opportunities for organisational actors to not only embark on a self-actualisation journey on their own but also to build a consensus upon which all individuals

are able to look critically on themselves and unpack the normative power/knowledge discourses (Crane et al., 2008).

Whilst, for instance, Habermasian application of ethics in business and organisations lies in institutionalising flat organisational cultures, democratic decision-making, and controlling corporate power (Flyvbjerg, 1998), Foucault's ethics resides in how organisational players [targeted by power] are able to set themselves free by acting differently (Alakavuklar and Alamgir, 2018). Thus, it is fair to argue that the Foucauldian theme of ethics formulates a new micro-emancipation pathway in the existing organisational cultures which has been driven by neoliberalism practices and corporate strategies (Munro, 2014). It was argued earlier how the recent breakthroughs in organisational processes due to the rise of AI/algorithms have led to many challenges affecting organisational stakeholders (Tsamados et al., 2022; Jarrahi et al., 2021; Vassilopoulou et al., 2022). In addition, we have seen radical changes in organisational structures due to the rise of new technologies such as algorithmic governmentality (Weiskopf and Hansen, 2022), biased selection and decision-making practices (Leicht-Deobald et al., 2019), workforce surveillance (De Vaujany et al., 2021) and more paramount issues such as the decay of human agency, dignity, and self-determination (Floridi and Taddeo, 2016). Algorithmic programmes have become the new circuits of power/knowledge in organisations (Clegg et al., 2006b) enacting new disciplinary work standards, strategically enforcing workforce subordination, whilst creating new regimes of truth by contextually changing the work practices (De Laat, 2019).

However, as literature highlights, the human end-users have been able to express their loathing and mistrust of their intelligent computerised work colleagues (Dietvorst et al., 2015) and resist their utilisation (Newlands, 2021). I explained how such resistance can fit and reconcile with Foucauldian problematisation of resistance and human subjectivity. It was discussed that the existing research and their theoretical cornerstones lack in-depth scrutiny around the discourses of resistance with several calls for research on this matter (Velkova and Kaun, 2021; Newlands, 2021; Anteby and Chan, 2018). Indeed, organisational scholarship has utilised post-structuralist philosophies to investigate the impacts of algorithmic tools on subjects and organisational discourses (e.g., Newlands, 2021; Lange et al., 2019; Bakir, 2015; Neyland, 2015). However, the possibilities for transformation of subjects [algorithm users, etc.], their self-awareness of algorithmic disciplinary mechanisms, and the development of ethical subjects have been theoretically overlooked or unexplored by the mentioned theories. The ethical dimension in Foucault's work, however, offers a critical insight through which not only can one explore how algorithmic work technologies shape new work control/dominative discourses (Weiskopf and Hansen, 2022), but also how organisational players articulate their subjectivity and agency within algorithmic work

regimes. It is indeed crucial to highlight again that Foucault never advocated the idea of ontological dualism of the self (Caldwell, 2007) nor ever posited any methodological approach for theorisation of the issue of agency and subjectivity (Fairclough, 1993). Yet, as it has been widely illustrated in the edited book by McKinlay and Starkey (1998), Foucauldian philosophical themes [especially the theme of disciplinary power/knowledge, surveillance, and control] have been quite influential in exploring post-Taylorism and post-Fordist types of organisations. Hence, it seems timely to put those aspects of Foucault's philosophy into work that are hitherto under-represented in the existing organisational scholarship (Crane et al., 2008) [such as the space for resistance or the ethical subject discourses], specifically for a debatable topic such as algorithmic work commodities and their ethical issues (Baum, 2020).

3.7 Conclusion

This chapter has drawn upon the analytical work of Foucault in order to sketch out a theoretical framework for this research. A Foucauldian approach through which the researcher understands and explores the discourses around algorithmic technologies and their ethical challenges in work practices.

The use of a Foucauldian theoretical lens entailed the theory of [critical] discourses that influence the adoption of algorithm tools as well as the novel disciplinary power/knowledge relations of algorithmic work regimes or algorithmic governmentality of the workplace. The theoretical landscape also touched upon the process of subjectification, activism of individuals, and resistance possibilities against algorithm tools. Foucauldian action and resistance aim to discover those discourses that oppose the existence of algorithmic risk assessment tools and/or voice the ethical concerns and issues around the utilisation of those tools. Consequently, I highlighted the latent pathway towards the development of Foucauldian ethical subjects. Although – as the literature indicted – this aspect of Foucault's work is ambiguous and incomplete, the emergence of algorithmic work cultures and their ubiquitous moral/ethical challenges triggers my interest to look into how Foucault's ethics theory will respond to the algorithmic issues. Not only that, but the topic of ethics in Foucault's work is generally underexplored in organisational scholarship. Thus, the theoretical void in the literature regarding the ethical discourses/issues of algorithmic work provides an excellent opportunity to tease out those discourses that can influence the development of Foucauldian ethical subjects.

These four key dimensions of Foucault's philosophical viewpoints are integrated to devise this research's theoretical framework. In the next chapter, I outline the methodology

underpinning this research to explore the perspective of key organisational stakeholders with regards to algorithmic risk prediction practices and their ethical issues. The next chapter indeed elaborates further on the researcher's philosophical standpoint [particularly the ontological and epistemological aspects] that would justify the use of Foucauldian theory for data analysis through the method of critical discourse.

CHAPTER 4: Methodology

4.1 Introduction

In the previous chapters, I examined the existing literature around the ethical side of utilisation of algorithms and explained the theoretical lens structured by Foucauldian philosophies. This chapter underpins this research's methodological approach. In recent years there has been a significant amount of research focusing on the ethical dimensions of AI and algorithm technologies, which are predominantly published in applied science academic outlets such as computer/data (e.g., Tsamados et al., 2022; Li et al., 2022; Mittelstadt et al., 2016; Kraemer et al., 2010). This array of research has significantly elevated our understanding around issues such as the value-laden persona of algorithms that can lead to bias or prejudices (Charlwood and Guenole, 2022). Indeed, such research has substantially contributed to crafting toolkits that help to ensure algorithms are used ethically (e.g., Prem 2023). These toolkits also aim to ensure that potential ethical issues such as biased decision-making are minimised by algorithmic decision-making (Floridi and Sanders, 2002). However, there are a few issues that I have identified that justify the selected methodological approach of this thesis. Firstly, the majority of the studies have considered ethics as an issue related to the technology itself. Such studies argue that ethical aspects can/should be fixed through appropriate supervision and stewardship by both human creators and users (Ouchchy et al., 2020). Therefore, human behavioural aspects such as the decay of human agency (Peeters, 2020) and resistances against algorithmic commodities are categorically overlooked. Second, the arguments around ethics are predominantly put forward by computer/data scholarship and tend not to properly address contextualisation within organisational environments. As such, a deeper understanding of ethics in algorithmic work practice will not only provide empirical evidence for specific ethical dimensions of algorithms but also extend existing knowledge beyond the domain of computer science. Philosophically, I will demonstrate how the ethics of algorithmic work practice can be applied as multiple 'truths' through constructivism.

4.2 The development of research questions

Silverman (2009) argues that research questions are pivotal in any research because they outline the directions the researcher should take to find relevant new ideas and appropriate data. Previously, I have outlined that the aim of this thesis is to critically understand the ethical dimensions of using algorithms in work practices. Also, I seek to better understand the power relations in algorithmic work milieux, which are areas where there is a paucity of

empirical organisational evidence. The two research questions that were developed to address these mentioned gaps, and have steered the research are:

- What are the dominant ethical discourses around the deployment of algorithms from the perspective of organisational actors?
- How do organisational actors influence the ethical discourses of algorithm tools and change their working experiences with these tools?

These two research questions are exploratory and take into consideration the perspective of key organisational stakeholders. Moreover, the theoretical perspective of this research is drawn on Foucauldian philosophy which seeks a better understanding of multiple discourses in different social context (Olssen, 2003) including organisations. Therefore, in order to justify these research questions, I explain, in the next section, why the choice of constructivism-interpretivism was deemed most appropriate.

4.3 Philosophical discussion and Research paradigms

The two key underlying factors that influence researchers' choices are *ontology,* which is how human beings perceive social reality, as well as *epistemology,* which defines what should be accepted as appropriate knowledge (Bryman, 2016). Reflecting on these philosophical assumptions establishes a crucial awareness that not only guides research projects but also indicates the wider impacts research outputs can have on social communities (Saunders et al., 2015b). Therefore, researchers may wish to reflect and align themselves with an appropriate philosophical standpoint that can help them with their investigation (Lincoln and Guba, 1985). I now explain my philosophical standpoint for this research and how it is aligned with constructivism (Denzin and Lincoln, 2011).

This study aligns itself with a position that understands and explores the social reality through subjective experiences (Burrell and Morgan, 2017). As such, I aim to understand how key organisational actors make sense of the social phenomena (i.e., their perspectives around the ethics of algorithmic work practices) and the meanings they attach to different dimensions of their social lives. The selected constructivist paradigm argues that reality is produced by social actors via a continuum of their lived experiences. It is the social actors who construct dynamics, context-specific and localised understandings of the events and phenomena (Schwandt, 1994). Moreover, constructivism suggests that there are no absolute 'true' social constructions, and they can differ in terms of complexity (Guba and Lincoln, 1994). Hence, my philosophical view fundamentally differs from a positivist paradigm, which

explains that the social world manifests itself in objective terms, including observable facts within enacted generalisations (Bryman, 2016).

The epistemological dimension is linked to the question of what is, and should be, considered as appropriate knowledge (Guba and Lincoln, 1994). In that regard, social events exist because the social actors inhabit them and give meaning to them (Bryman, 2016). For the purpose of this research, I see the social reality as a product of the meanings given to it by its inhabitants. In this research context, participants and the researcher are the social actors, which grounds this research in 'interpretivism' (Blaikie, 2009). Epistemology also entails methods to observe the social world and make sense of it. Producing knowledge demands a level of understanding around what that knowledge includes (Crotty, 1998). In a similar manner, Cohen et al. (2013) argue that epistemological assumptions are those around the bases of knowledge, how it is shaped, and how it can be understood and illustrated. Such discussion is also in line with the adoption of Foucault's theories. Foucault's earlier work as a historian involves his views of dominant discourses, which informed his later work; though, he never devised a particular methodological approach (Graham, 2010; Heller, 1996). Yet he argues that discourses are mediators through which social actors produce their own knowledge and form 'control over things' (Fairclough, 2003). As such, Foucault was epistemologically concerned with how discipline is shaped and how people become self-disciplined through power/knowledge (Knights, 2002).

In line with this, I argue that what is narrated by the social actors and the researcher's own interpretations shape the necessary pillars of produced knowledge (Bell et al., 2018). As an interpretivist, I believe that my own account of a social phenomenon is a construction in itself, hence, advocating the idea that I am not presenting or reporting a definite view of reality but rather my own version of it (Bryman, 2016). Furthermore, I acknowledge that as an interpretivist, the social reality is not objectively measurable, but it is subjective and interpreted by the participants of the research (Lee and Lings, 2008).

In terms of the relationship between theory and the research strategy, I align myself with *abductive reasoning* (Saunders et al., 2015b). Abductive reasoning is an alternative approach to deductive and inductive approaches for theory usage (Clark et al., 2021). A deductive theory development necessitates testing hypotheses by the use of empirical data, which results in confirmation or rejection of those hypotheses. The goal of a study in a deductive approach is to revise existing theory (Bryman, 2016). In contrast, an inductive approach favours building theory from scratch through findings and observations. Put differently, inductive research generates theoretical significance out of empirical findings (Strauss and Corbin, 1990). Abductive reasoning, however, considers both of the mentioned

approaches as it switches back and forth from observations/findings to the existing literature and social world. In abductive reasoning, the researcher thinks about data and theory at the same time and acknowledges that there are no certain conclusions, but only plausible ones (Flick, 2013).

Foucauldian theories, which ground the foundation of this research, are essentially concerned with the organisation and administration of people (Bergström and Knights, 2006) and the extent to which people's discourses shape the social world (Fairclough, 2013). An abductive approach can be aligned with Foucauldian theories as it considers people and their perspectives at its core (Charmaz, 2014). According to Clark et al. (2021) an imperative step in abductive reasoning is to see the social world through the eyes of people being studied and understand their discursive practices, language, meanings, and viewpoints in the study of the social world. Hence, I have subscribed to abductive reasoning because, firstly, it connects the findings and existing theory in a *shuttling* way (Atkinson et al., 2004), which impedes the researcher from making unsubstantiated inferences. And secondly, because the abductive approach is well aligned with the Foucauldian lens of this research, as they both aim to understand the social reality through the eyes of their participants and the multiple discourses that people adhere to.

Exploring ethics in relation to algorithmic technologies in work environments encompasses the interpretation, perspectives, and discourses of key organisational stakeholders. This exploration requires an actor-centred understanding in a context where transformations take place due to the introduction of algorithms. The everyday interactions, work experiences, discourses, and artefacts of social actors in relation to algorithmic agents create a social setting that is peculiar and is continuously reproduced (Kivunja and Kuyini, 2017). Therefore, exploration of ethics is interwoven with the work experience of actors and how the key actors (including the researcher) interpret these relevant social interactions. Although interpretivism is not exclusively associated with *qualitative* research (Ritchie et al., 2013), many interpretivist researchers are typically inclined to choose qualitative research strategies to undertake their projects. The next section explains the qualitative design used to operationalise this research project.

4.4 Research Strategy and the Choice of Qualitative Strategy

Qualitative research is frequently derived from a constructivist approach, which entails iterative engagement with empirical evidence to shape a picture of social reality and generate theoretical contributions (Bryman, 2016). The adoption of qualitative inquiry was considered most appropriate for two reasons. First, from an interpretivist standpoint, I was

keen to illustrate the social actors' interactions and work experiences with algorithms (Bryman, 2003) that shape ethical dimensions of algorithmic work practices. In other words, grasping ethics in algorithmic work practice through the perspective of social actors was an integral aspect in my research. Furthermore, as I explained in the literature review chapter, there are underexplored areas in relation to the ethics of algorithmic work practices, such as the locus of power and space for resistance. Power, agency, and resistance are also burgeoning areas in the discussions around algorithm ethics that specifically concern the perspectives and interactions of organisational actors (De Laat, 2019; Weiskopf and Hansen, 2023). However, as I outlined previously, the relevant literature lacks rich empirical evidence around the mentioned topic, particularly in organisational settings. Hence, I aimed to explore and understand actors' experiences and perspectives in an organisational setting to uncover the ethics of algorithms through the research questions.

Second, Foucauldian theories such as discourse, power/knowledge, and governmentality are mostly centralised around people, their discourses, and/or discursive practices (Barratt, 2008). Therefore, by adopting a Foucauldian theoretical lens, this study aims to investigate how organisational actors working lives are transformed through algorithmic work practices. Thus, in this research, I do not intend to *test* any proposed ideas through the collected empirical evidence, which is aligned with positivist thinking. Rather, I sought to gain a deeper understanding – through a Foucauldian lens – of how organisational players unfold ethical dimensions of algorithm ethics.

It is argued by Creswell and Poth (2016) that when there is a need for a deep, rich, and contextualised understanding, the use of qualitative methods can be appropriate, as such methods help to explore and unpack nuances and gain a deeper understanding. Miles et al. (2018) discuss that not only the utilisation of qualitative techniques can "lead to serendipitous findings and interrelationships," but also, they help the researchers to "generate new understandings" (p. 3). The research questions are developed to respond to the scarcities in the AI/algorithm ethics literature. Therefore, the use of qualitative methods can be effective to generate new dimensions of ethics, contributing to the relevant literature. Qualitative research techniques offer more flexibility and enable researchers to identify novel in-depth data compared to quantitative methods (Bryman, 2016). Similarly, Edmondson and Mcmanus (2007) discuss how incorporating qualitative methods such as interviews, observations, and documentation analysis provides profound and rich data for the studies with nascent theories. Therefore, I make use of qualitative data to unpack the ethical nuances of algorithmic tools in work practices and contribute to the existing knowledge.

Based on the insights from Yin (2018) undertaking a meaningful research project firstly requires rigorous planning, including the synthesis of the research problem, aims, and objectives. Following this is the design stage, which encompasses methodology, development of conceptual frameworks, designing questionnaires, etc. The next is the preparation phase, in which the researcher pilot tests (pre-tests) the questionnaires and interview schedules. This is followed by the data collection phase, which includes the interviews and transcription procedure. The analysis phase consists of reducing and/or condensing the data, such as the coding process by e.g., CAQDAS, such as the NVivo ® package. During this stage, the researcher displays the data and subsequently finds patterns, contrasts, and causal flows in the data in order to draw conclusions. And finally, the findings and discussions of academic studies should be shared with the wider academic community. As such, to fulfil the final stage, empirical papers are to be published, and participating organisation(s) will receive a report of findings so they can enhance their algorithmic work practices. I will now explain each stage in more detail below.

4.4.1 Research Design and Setting

Research methods scholars such as Bell et al. (2018) have listed different research design categories such as case study, comparative case studies, survey study, experiment, and longitudinal study, which can be associated with qualitative inquiry. A case study, as a research design, is an empirical method to explore in-depth a phenomenon, i.e., the 'case' within a real-world context (Yin, 2018). Put differently, a case study, as a mode of enquiry, investigates a distinctive social situation thoroughly in a selected context (Bryman, 2016).

A case study research design was deemed most appropriate to address the research questions highlighted above. There are two justifications for my choice of case study research design. Firstly, I needed an organisational setting with specific features (Bryman, 2016), including algorithmic work environments and awareness of ethical implications. This research required a unique organisational case where algorithmic transformations and ethical issues distinguish that particular case from others and has made it suitable to address the research questions (Bell et al., 2018). Secondly, the case study design enabled an intensive examination of the complexities and contradictions of real-life phenomena (Yin, 2018). This provided an opportunity to identify theoretical and practical contributions, further strengthening the link between the theory and research. Furthermore, case study design is predominantly considered as informative for both inductive and abductive reasoning since it combines theories with empirical evidence (Hartley, 2004). Hence, I argue it is strongly suitable to carry out a case study as this research aims to look at the ethical dimensions of

algorithmic tools contextualised in work environments through a Foucauldian theoretical lens. And subsequently be able to provide further theoretical insight around the issue. Therefore, a case study research design was considered most suitable.

Initially, in order to select the most appropriate case study that addresses the research question, I reached out to several companies, firms and organisation in both public and private sectors operating overseas and here in the UK. These organisations were targeted because their working processes have been transformed by algorithmic tools or different AI technologies. For instance, I reached out to a potential company because of implementation of semi-autonomous robots in their supply chain practices. In the process of selecting case studies, I also got in touch with particular individuals within those organisations, briefly explaining my research rationale, aims and objectives. These communications were undertaken either via emails or letters.

A criminal justice organisation, based in a country in the global north, answered my call and expressed their interest to participate in this research. This organisation has introduced a range of algorithm tools for the purpose of risk assessment, offering suggestions on suitable rehabilitative programmes for offence cases. Following this first contact, I provided a more detailed version of the project overview sheet containing details such as research rationale and background, main research questions, research methods, as well as potential outcomes and benefits. I had an informal discussion with one of the directors, and they reiterated that they are aware of the ethical ramifications the algorithms may have on people and aim to resolve them. This particular criminal justice setting was considered most suitable as a case study because of a few reasons. First, the organisation's operations were at a national level, which means that the researcher could access a range of regional divisions, exploring different social relations and perspectives (Silverman, 2009). Second, the choice of criminal justice as the case study is justifiable because case studies are always theoretically guided (Yin, 2018). Criminal justice organisations can be particularly interesting 'research cases' since there are growing concerns amongst scholars around ethical issues in such organisations, including biased decision-making (Cunneen, 2006) or dystopian challenges such as 'systemic racism' (Davis, 1996). And such ethical concerns have been further intensified due to the emergence of AI and algorithmic decision making in criminal justice (Završnik, 2021; McKay, 2020). Furthermore, Foucault's notions, including discourse, governmentality, administration, organisation, and disciplining people (Bergström and Knights, 2006), are important theoretical concepts in this research. Some of these concepts are also reflected and contextualised in criminal justice organisations (Garland, 1997). Hence, the opportunity to undertake research in a criminal justice organisational setting in the form of a case study provides rich contextualised insights around ethics. Also,

due to the deployment of a range of algorithmic technologies, which has raised concerns around the ethical side of these technologies as well as time constraints of the PhD programme, I decided to undertake my research in this organisational setting. Access negotiations also took place via interviews with the executive directors of the organisation as well as the performance team panel, which evaluates research independently. Unfortunately, due to COVID-19 and a shortage of staff members, the ethical approval process took longer than expected.

It is worth mentioning that the researcher and the supervisory team had initially contemplated the idea of doing a multiple-case study design. This design entails research when the number of cases exceeds one (Bell et al., 2018). Bryman (2016) argues that comparative design helps for better theory building when there are meaningful contrasts between the cases. Yet, Clark et al. (2021) discuss that multi-case design can be problematic because of issues such as access, equality in the number of participants, or even funding. Due to the accessibility issues and the time constraints of the PhD project, a multiple-case study design was discarded.

4.4.2 Data collection phase

Qualitative research can incorporate a variety of methods to collect data and address research questions. Mason (2017) names four main techniques for collecting qualitative data, which are interviews, observations, documentation analysis, and visualisations. Creswell and Poth (2016) recommend that researchers collect information using multiple sources of data. This study, thus, has adopted two sources of evidence. I will outline the suitability and selection of these methods below.

4.4.2.1 In-depth interviews

For the purpose of this research, in-depth semi-structured qualitative interviewing was selected as the primary technique for data collection. For a case study research design, interviewing is a pivotal method for collecting rich information (Yin, 2018). Moreover, interviewing is deemed the most appropriate method since it is quite consistent with interpretivist epistemology, as highlighted above. Also, interviewing is considered as an effective route to access and explore people's mindsets, perspectives, and meanings, and how they define situations and construct reality (Punch, 2013). Through interviews, researchers can study each person's personal perception of social phenomena in detail and gain in-depth understanding of that particular context where the social phenomena have occurred (Ritchie et al., 2013). Interviewing was considered crucial for this study as the researcher was keen to explore the ethical aspects of algorithmic work practice through the

perspectives of key organisational stakeholders. Thus, it was necessary to explore how these organisational actors perceive ethics around algorithmic apparatuses and how their discourse transforms the work practices. The selection of interviews as the main methods of data collection has enabled the researcher to better probe people's standpoints to achieve the depth of answers from the individuals (Clark et al., 2021).

Additionally, in-depth interviewing was considered the most suitable technique of data collection as it provides the opportunity for clarification and probing (Ritchie et al., 2013). This is particularly important as the PhD researcher is not a native English speaker, and understanding specific expressions, metaphors, or idioms was sometimes challenging to grasp. Hence, I piloted the interview questions to identify and rectify any issues where required. Semi-structured qualitative interviews were the main method of collecting information, which enabled me to mitigate the latent language barriers. In a cross-country research project conducted by Mogaji and Nguyen (2022) on the developing AI for financial marketing service, qualitative interviewing was selected to ensure clarity and overcome any language barriers.

I selected semi-structured interviewing rather than the unstructured type since I needed to explore particular dimensions that are relevant to algorithm ethics. These ethical dimensions are revolving around people's perception of the process of adoption, the efficiency and/or flaws, as well as their own views on algorithms and the future of criminal justice organisations. Unstructured interviewing may not have captured all these momentous nuances, as it is much more flexible and very similar to a *conversation* (Bell et al., 2018). Furthermore, because of higher levels of flexibility in unstructured interviewing and – potentially – lack of an interview guide, it is difficult to establish a consistent approach to interviewing. In semi-structured interviewing, however, I was able to tailor my research questions, objectives, and points of interest in the form of an "*interview schedule*" (Clark et al., 2021, p. 426). Also, by using interview schedule, I was able to prompt the question, probe answers, establish better connection with my participants and ask the questions in a more open and empathetic manner (Hennink et al., 2020).

Although there are several advantages associated with qualitative interviewing, such as being less prone to unnatural reactions from participants or less intrusive in people's working lives compared to participant observation (Bryman, 2016), it is not without downsides. For instance, Silverman (2009) discusses that when people are invited to tell their story of past behaviour, they are positioned as 'authentic', which promotes an overly rationalistic perspective of human behaviour. This might result in concealing the true identities of individuals and only presenting words in isolation. Bell et al. (2018) argue that

due to heavy reliance of interview on verbal accounts of behaviour, implicit features of social interactions are less likely to surface. However, I believe that due to the particular focus of this research, which is ethics in algorithmic work contexts, the merits of interviewing outweigh its disadvantages. As it is argued by Bryman (2016), “qualitative interviewing enables the researchers to maintain a specific focus” (p. 486) throughout the whole research.

It is suggested by Bell et al. (2018) that qualitative researchers should be strategic in terms of sampling. *Probability* sampling, which is predominantly associated with quantitative research, may not be the best option for some qualitative research. This is because, unlike quantitative design, generalisability and representation of the population are not usually amongst the key criteria of most qualitative work (Bryman, 2016). Unlike quantitative studies, qualitative research is more concerned with the nuances of particular social phenomena by engendering an in-depth understanding of the social actors’ perspectives, which can have theoretical significance (Silverman, 2009). Therefore, rather than random probabilistic sampling, I have used *purposive sampling* (Clark et al., 2021), of which I have targeted particular individuals across the organisation who were most relevant to the research questions of this study. In order to conduct a robust sampling, I have subscribed to the guidelines by Bell et al. (2018) and applied the *snowball sampling* technique to approach the participants. In order to carry out snowball sampling, I initially contacted a small group of employees in the targeted organisation who were most relevant to my research focus and used them as contact mediators to get in touch with other relevant organisational members. The first interviewees were asked to suggest other potential members who were either directly involved in the processes of ARA adoption or have working experience with the tools. According to Bryman (2016), it is vital for the researchers to conduct rigorous assessments about the recommended interviewees. The researchers should make judgements on whether a participant should be included in the study, rather than solely relying on the recommendations of the initial participants. Therefore, in order to ensure that the characteristics of recommended participants match the aims of this research, I was provided with email addresses and had some communications to understand if they were willing to and are suitable to be included in the study. To do so, I provided the potential interviewees with the project brief and explained the research aims and objectives via back-and-forth email communications. In these communications, I also explained the ethical implications of interviewing, including voluntary participation rights, confidentiality, and security of their data. These initial communications helped me to ensure that the recommended participants’ knowledge and/or experiences are aligned with this study’s goals.

It is also crucial to highlight that email communications and the recruitment process of particular interviewees (i.e., the frontline practitioners) took longer than originally planned: By February 2023, I had only conducted three interviews with frontline practitioners. Later, I was informed by the organisation that these practitioners might have been reluctant to participate due to their significant workload and time constraints related to clients processing. After careful consideration of the issue and further discussions with the supervisory team, I decided to encourage these practitioners by offering financial incentives (Head, 2009). To do so, I added a note in the email communications stating that each interview participant would receive a gift card (£10 Amazon voucher) as a gesture of my gratitude for agreeing to take part in my research. The funds for the gift card were kindly offered by my external supervisor sourced from their available research funds at Aston University. This change was also highlighted in the research ethics application, communicated, and approved by Aston University Business School Ethics Committee. Incorporating a payment strategy in the requirement process proved fruitful as I managed to receive many expressions of interest for interviews over the course of a few weeks.

In qualitative interviewing, it is almost impossible to anticipate how many people should be interviewed so it is deemed adequate (Bell et al., 2018). In other words, proposing any number as an appropriate sample size would be arbitrary in qualitative interviewing. For this purpose, Charmaz (2014) argues that the principle in qualitative interviewing should be continued until data no longer indicates new theoretical insights or dimensions. Theoretical saturation is the term used for the mentioned principle (Clark et al., 2021), in which I continued to sample more interviewees concurrent with preliminary analysis of data. I carried on with sampling up to the point where a complete theoretical understanding was achieved. To simplify, interview sampling was halted as soon as I inferred that themes emerging from the data had become repetitious.

Furthermore, semi-structured interviewing is better aligned to address the research questions, as there is relatively a clear area of focus [unit of analysis] with regards to understanding ‘ethics’ of algorithms. Semi-structured interviewing is preferable by the research as it ensures consistency and balance due to utilisation of an ‘interview guide’ (Bell et al., 2018). An interview guide usually contains identified issues that a researcher aims to address. An interview guide is preprepared for interviewees of a research and normally includes a list of questions drawn from relevant literature. Essentially, an interview guide helps the researcher to uncover the most useful information for answering the main research questions (Lofland et al., 2022). As such, I provided each interviewee with an interview guide in advance of the interviews. I crafted different types of interview guides, depending on the

roles and job descriptions of the interviewees. For more details regarding the interview guides, please refer to Appendix 4.

All interviews were carried out using the Microsoft Teams app, which is being used by the targeted criminal justice organisation as the main online meeting platform. Using Teams has several advantages for both the researcher and the interviewees. First, the process of data collection from June 2022 until June 2023 roughly overlapped with the end of the COVID-19 pandemic and remote working fashion flexibilities. Subsequently, many employees in that organisation have adopted flexible working patterns. Thus, using Teams allowed me to better schedule interview appointments and conduct them in a more flexible manner that suited the interviewees. Second, Teams provides features such as recording and auto transcription. These features allowed me to record the interviews with just a click of a button and have transcripts ready in a Word document. That being said, the transcriptions generated by Teams were not totally accurate and still required editing. Yet, regardless of the need for editing, auto-generated transcriptions significantly streamlined the interviews and enabled me to allocate more time for the analysis stage. Please refer to Appendix 7 for a sample of generated transcripts via MS Teams. This sample is edited and anonymised.

Conducting interviews via online platforms, however, was not without downsides. For instance, I was not able to fully observe the facial cues or body language of my participants, especially during a few interviews where the camera was switched off by the interviewee. Furthermore, on these few occasions, I felt that the rapport between me – who is a stranger – and the participant was not fully established. Thereby, it seems that the participant may be reluctant to fully expand on the topic. In order to overcome these challenges, I used more warm-up questions with those participants with the aim of gaining their trust. Also, in a few interviews, I encountered some technical problems due to hardware breakdown (e.g., battery drainage) or internet connection (intermittent connection). I was able to rectify these issues by quickly switching to a second computer or using mobile hotspot internet. Table 1 illustrates the details of the interviewed organisational members.

Table 4.1: Details of interviewees.

	IDENTIFIER	Gender	Role specification
Senior management [n=3]	SM1	M	Director
	SM2	M	Deputy-Director
	SM3	M	Forensic Psychologist
Data Science Team [n=9]	DS1	M	Head of Data Division
	DS2	F	Head- Data Engineering
	DS3	M	Chief Data Scientist
	DS4	M	Chief Data Scientist
	DS5	M	Chief Data Scientist
	DS6	M	Head-Data Linking
	DS7	F	Data Ethicist
	DS8	F	Head-Corporate Data Science
	DS9	F	Chief Data Scientist
Human Resource Manager [n=2]	HRM1	F	Deputy director-workforce experience
	HRM2	M	Divisional director
Line manager (senior frontline) [n=7]	LM1	M	Regional team leader
	LM2	F	Regional team leader
	LM3	F	Regional team leader
	LM4	F	Regional team leader
	LM5	M	Regional team leader
	LM6	M	Regional team leader
	LM7	F	Regional team leader
Frontline practitioner [n=17]	FP1	F	Frontline practitioner
	FP2	M	Frontline practitioner
	FP3	M	Frontline practitioner
	FP4	F	Frontline practitioner
	FP5	M	Frontline practitioner
	FP6	F	Frontline practitioner
	FP7	F	Frontline practitioner
	FP8	M	Frontline practitioner
	FP9	F	Frontline practitioner
	FP10	F	Frontline practitioner
	FP11	F	Frontline practitioner
	FP12	F	Frontline practitioner
	FP13	F	Frontline practitioner
	FP14	F	Frontline practitioner
	FP15	F	Frontline practitioner
	FP16	F	Frontline practitioner
	FP17	M	Frontline practitioner
Total:	38		

4.4.2.2 Documentary data

The second source of information was documents published by the organisation on their algorithmic practices. Yin (2018) argues that documents play a pivotal role in corroborating and augmenting the evidence collected from other sources such as interviews or ethnographical observations. As such, documents such as an organisation's publications are often articulated as complimentary secondary data that can strengthen qualitative studies (Denzin and Lincoln, 2011). Following this, I carried out searches and managed to access a number of documents relevant to the research focus of this study and were informing the ethical side of algorithmic work practices. These documents were accessible to the public around the internet and predominantly devised by the data science division. These documentations used ranged from the organisation's digitalisation strategies to user guidelines for ARA tools and ethical frameworks of ARA practices.

It is important to highlight that the documentations used were not treated as the main source of data, but they were rather additional data sources alongside the qualitative interviews (Bell et al., 2018). This is because the published documents predominantly incorporated the views and perspectives of only a small group of key organisational actors, i.e., the senior leaders and data scientists. The accessed documents, therefore, lack inputs from the wider community of frontline practitioners who are the end-users of algorithmic tools. Moreover, as discussed by Bell et al. (2018), organisational documents accessible to the public might be written in a way to "promote a favourable view of the organisation to outsiders" [p.532], including external researchers. It suggests the authors of organisational documents (e.g., the data scientists in this research setting) are likely to have a particular discourse that they wish to get across. Thus, it was evident that the documents used for this research, on the surface, may offer only limited representativeness of key organisational issues or failings, and drawing insights from such documents may not capture the points of view of key frontline professionals who may have an important 'say' with regards to the ethics of using algorithmic technologies. However, from a Foucauldian perspective, documents could also be seen as the main external way of promoting the dominant discourse.

Table 4.2: A summary of collected data

Source	Example	Source Format	Details	Notes
Primary source (created by the researcher during the case study research)				
Online Interviews	Interview with Head of Data Science at the Criminal Justice Service case study	.docx transcripts	***38*** one-to-one semi-structured interviews with key organisational members	Generated by the speech recognition tool via MS Teams; Edited and refined by the researcher
Additional documentary source (generated and published by the organisation)				
Reports	A compendium of research on the use of ARA systems	Documents (PDF)	***4*** (Accessible for public)	Social research and statistical analysis conducted internally by the organisation's management service
User Guidelines	A guideline for practitioners to use ARA for predicting risk of serious offence.	Documents (PDF) & (.docx)	***4*** (Accessible for public)	Devised and heavily influenced by the creators of ARA tools: The Data Science division
Strategic plans	Reducing Reoffending plan	Documents (PDF)	***2*** (Accessible for public)	Collaboration between the criminal justice service and independent institutes to develop ethical frameworks (toolkits) for AI and algorithms

All in all, this study gains benefits from two sources of data: semi-structured interviewing as the main source and documents as an additional source. Further details on the two sources of data used for this research can be found in Table 2. The above section has explained the techniques that were chosen to collect data as well as the reasons for choosing these techniques. The next section elaborates on the ethical considerations associated with this research.

4.4.3 Ethical Considerations

Ethical issues are amongst the most crucial aspects of any research project. And it is important for the researchers to consider the ethical issues in early planning stages (Robson and McCartan, 2016). All researchers working in the affiliated university are required to evaluate the ethical issues associated with their research. They are asked to conform to the University's ethical guidelines and comply with the Research Ethics Framework of the ESRC. Subsequently, they are required to submit their applications to the Ethics Committee at the University Business School for approval. The proposal and ethics application should be approved prior to any data collection taking place, according to the University's ethics framework. According to Silverman (2009), there are a couple of ethical issues that could arise in doing research and should be mitigated by the researchers. Firstly, it is the matter of informed consent. To ensure this, the researcher aimed to ensure all participants are provided with the details about the research either directly through their emails or by the key gatekeeper [point of contact]. To do this, the researcher produced a participant information sheet and a consent form to be sent to all participants. These documents entailed key information around the purpose of the research, the way the data will be used, and what will be required from the participants. All participants were asked to carefully read the information sheet before giving their consent. I also encouraged the participants to ask any questions regarding the study prior to giving their consent. Furthermore, the informed consent included the voluntary withdrawal rights reserved for both the organisation as well as the individuals. Direct contact information, including my email address and mobile number, was provided for the organisation and the participants in case they had any questions.

Secondly, it concerns the anonymity and confidentiality of the participating individuals. To ensure this, all interviewees were informed that any information that might reveal their identities will be removed or changed from the research, and no participants will be identifiable in the research outputs such as the PhD thesis, conference, or journal papers. Moreover, I aim to maximise the anonymity and confidentiality of the interviewees by

removing any personal details from the transcripts and assigning identifiers to the participants instead of their names (Saunders et al., 2015a) [e.g., FP4 for frontline Practitioner 4, DS2 for Data Scientist 2 or LM6 for Line Manager 6 and OD for Organisational Document]. Also, I changed the name and context of this research's case study to criminal justice service in a global north [European] country to avoid revealing the country or cities where the research took place. That said, it is impossible to achieve total anonymity since in any research there are always parties who are aware of the true identity of the participants or case study organisations, such as the research team or the candidate's supervisors.

And finally, the researcher aimed to protect the participant [including the organisation] from harm. Silverman (2009) argues that the social research process should be carried out in a way to minimise the risk to the individuals in the research. This is particularly vital to ensure that participants' interests or well-being would not be damaged as the result of participating in the research. In order to ensure this ethical dimension, I outlined that this research is only exploratory and highlighted to my participants that as an external researcher, I am not in a position to change the organisation's policies or practices. By reiterating this in the participant information sheet, I aimed to ensure that I would not be conveying any misleading message or hopes to the participants. Additionally, the interview questions were evaluated by me, the supervisory team, and the Aston Ethics Committee in order to ensure that they don't pose any psychological discomfort to the participant. Also, at the beginning of each interview, I made the participants aware that at any point during the interview they can ask me for a break or to stop.

The above points were the main principles that I have invoked to ensure the research process is conducted in an ethical way. There are, however, other ethical considerations which that be explained in more detail. Data management is an additional ethical issue associated with participants' confidentiality. It is essential to ensure the data collected from the participants is protected and secured throughout the research, including storage or dissemination and publications (Bell et al., 2018). For this purpose, I have complied with EU GDPR 2018 act guidelines and followed the University's advice on data protection issues and stored all personal information on the Box app. The Box app is a cloud storage approved by the Ethics Committee and academics at the affiliated University. The anonymised data will be kept for no longer than 10 years.

Clark et al. (2021) highlight funding and conflicts of interest as ethical concerns that could affect the researcher's independence. These issues can discredit a study and deem it as biased. Therefore, researchers should be transparent and explicit about the sources of

funds and support (Bell et al., 2018). This research project was self-funded and carried out solely by the lead researcher with no conflicts of interest. The next section examines the data analysis procedure of this research.

4.5 Data analysis

The collection of data highlighted above resulted in approximately 32 hours of audio recordings. The data included 1,212 pages of textual information, including 683 pages of interview transcript and 529 pages of organisational documents. In this section, I explain how the Foucauldian Discourse Analysis is used to analyse the data and draw meanings from it.

4.5.1 Foucauldian Discourse Analysis (FDA)

In his book, *The Archaeology of Knowledge,* Foucault raises an awareness of the notion that language and discourses have materiality and therefore are capable of producing ideas manifested in social practices (Foucault, 1972). Moreover, he unpacks how discourses are instruments for power exercise and enable the people to constitute subjectivity in different socio-cultural contexts (Bryman, 2016).

According to Graham (2011) defining discourse depends on one's epistemological approach to social phenomena. For instance, for Phillips and Hardy (2002), discourse is an interwoven combination of texts, their production and dissemination that makes a thing into a being. That is to say, social reality, including the interactions, conventions, and artefacts, is all produced and made real through discourses. Keller (2011) conceptualises a discourse as a frozen state of meanings in time that institutionalises between actions and agency within social collectives. For social theorists such as Van Dijk (1985), Fairclough (2013), and Foucault, the premise of power is a momentous component. The Critical Discourse theorists, such as Van Dijk or Fairclough aim not to substitute one "truth" for another. In other words, they acknowledge that there can never be one universal truth or absolute ethical standpoint. Foucault, however, interprets discourses slightly differently. According to Foucault, a discourse is a powerful way of producing "things" (Graham, 2011). For him, words, statements, or sentences are not just a way of utterance but *functions,* which can be shaped or deployed to constitute social formations or actions (Keller, 2011). Also, Foucault explains in his book, The Archaeology of Knowledge, that discourses are systems that form knowledge(s).

In that regard, he explores different examples of discourses such as psychiatric discourse, natural history discourse and the discourse of clinic (Foucault, 1972). Indeed, by thinking around the premise of knowledge and its relation to discourse, Foucault has moved

away from the linguistic tendency to discourse [which is predominantly associated with the term 'discourse'] and towards 'systems of dispersions.' Systems of dispersion, as explained by Foucault, are rules, structures, formations, and thematic choices that are written or unwritten and that form a set of practices (Mills, 2003). Hence, the way Foucault uses the word 'discourse' is not related to language, utterance, or communications in social collectives but refers to *the way we do things* (Foucault, 1972). According to Foucault's discourse theory, people speak, think, or write about a particular social object only in particular ways and no other. Thus, a 'discourse' is that instrument that enables people to think, speak, or write about a given social object, but also what constrains it (McHoul and Grace, 1998).

For Foucault, power is associated with discourse and inscribed into it. According to McHoul et al. (2015), discourse is the thing that depicts power relations or struggles and does not mask it. In other words, discourse is the power that should be appropriated. As such, one can argue that Foucauldian theories such as power/knowledge, governmentality, subjectification, and resistance are all crystallised via discourses within different social interactions (O'Farrell, 2005). Foucault's concept of 'power' in Foucauldian discourse rejects power as only top-down monolithic repressive dynamic that aims for "production, giving rise to new behaviour" (Mills, 2003, p. 33); rather, power can be exercised within everyday interactions and relationships between people and institutions. Foucault explained the process of 'subjectification' through which subjects' discourses can face a transformational shift (Heller, 1996). The subjectification process – which is transformation power discourses – occurs when objectified voices challenge any dominant discourse (power) that is being exercised over them. In that sense, Foucault believed that any discourse can be reversed; specifically, objectified and marginalised discourses can inform new alternative views or knowledges (Fairclough, 2003).

For the purpose of this study, I subscribed to FDA theory and applied it to explore different discourses that can be linked to 'knowledge' around ethics of algorithmic work practices. I believe that FDA theory is particularly informative since it considers power relations, the production of knowledge, and the subject's agency. And as it was discussed in the literature review chapter, there are underexplored areas in relation to the ethical side of algorithmic work practices and it requires inputs from key organisational actors. Thereby, FDA theory seems a relevant method to uncover those underexplored ethical notions.

It is imperative to highlight that Foucault never suggested any particular model, protocol, or guidelines to conduct discourse analysis (Graham, 2011), and it is indeed challenging to apply Foucault's rather abstract ideas to any research (Arribas-Ayllon and

Walkerdine, 2017). Yet, Foucault (2010), in his later lectures, explains a project to understand how contemporary practices (discourses) are being used by people to constitute themselves as subjects of knowledge. He highlights that understanding 'discourse' is to understand *subjectivity* and *experience* along three correlated axes (Foucault, 2010): First, is the axis of *knowledge,* which he discusses in his archaeological work (Foucault, 1972), and re-theorises that discourses are not solely instances of text or utterance but systems, rules, and/or rationalities within a specific body of knowledge (Fairclough, 1993). He highlights that such rules or systems are *governing* discursive practices; governing rules or strategies that can outline what is true or false in different types of knowledge, such as psychiatry, medicine, and judicial, etc. (Hardy and Thomas, 2015).

Second is the axis of *power,* alluding to his genealogical work on power/knowledge (Foucault, 1977). This axis is related to understanding how behaviours and/or conducts of people can be controlled or governed by particular discourses or discursive rationalities (Townley, 1993). Such discourses, as Foucault argues, are developed to ensure subjects' self-governance and normalisation (Clegg et al., 2006a).

And finally, it is the axis of *ethics,* which is related to his later work on the Roman and Greek ethics and subjectivity (Foucault, 1990; 2019). In that regard, Foucault considers discourse as 'positions' that subjects take to act against 'subjection' and working through acts of 'subjectification' (Arribas-Ayllon and Walkerdine, 2017). Subjectification, for Foucault, means the discursive processes through which individuals problematise power and transform themselves to gain a particular state of awareness, morality, and/or perfection (Barratt, 2008). This axis, therefore, is related to practices through which people constitute themselves as 'ethical subjects' by being aware of power dominance (Skinner, 2013). In relation to the focus of this research, 'ethics' in algorithmic work, I look for 'discourses that underscore expressions, phrases, contingent rules within any discursive pattern, (un)written utterance, or statements that are meaningful (Foucault, 1972). By 'meaningful' I do not intend to dig up the intangible meanings within the language, but to understand the rules, formations, and systems within utterances/statements that can constitute particular 'algorithm ethics' or may control and delimit other discourses.

To do this, I have followed Fairclough's (1993) suggestion and problematised how the ethics of algorithms as 'objects of discourse' are constructed in a specific setting, such as for organisational work practices. In other words, problematising ethics in relation to algorithmic work practices foregrounds the context in which I can think differently about the existing regimes of truth, or the established knowledge of AI/algorithm ethics.

Another aspect of the FDA relevant to this research is “Technologies of the Self” (Foucault, 2020). By this, Foucault refers to systems or assemblages that act on human conduct from a distance. This term also refers to discursive practices of power through which people are able to constitute themselves as subjects. In light of algorithms at work, this is reflected within the discourses of organisational elites who advocate the adoption of algorithms (Pachidi et al., 2021) as well as the examples of resistance against algorithm work regimes (Anteby and Chan, 2018).

Overall, I reiterate that FDA is a way to examine discourses by considering specific notions such as power. Therefore, the FDA provides no specific framework on *how* to analyse text, statements, or documents but provides an analytical lens to examine discursive patterns in relation to, e.g., social power relations and their relevant discourses. However, FDA as a form of discourse analysis still requires the researchers’ rigour in analysing interpretations and explorations (van Dijk, 1985). Also, there is a tendency amongst discourse analysis researchers to look for reoccurring patterns in the textualized materials. In that regard, reflexive thematic coding (Braun et al., 2022) was an effective method to be combined with FDA in order to analyse the data in-depth. The reflexive thematic method emphasises researcher reflexivity in the sense that themes should not be considered as pre-existing codes awaiting retrieval; themes are indeed not ‘obvious’, but they are discursive patterns with central meaning-based concepts (Braun and Clarke, 2019). I developed the codes with deep reflections in order not to treat them as solid codes, but as ‘organic interpretive stories’ produced through the intersection of the researcher’s theoretical reflections and the data itself. Braun and Clarke’s (2019) argument on reflexive thematic research enabled that and is aligned with what Foucault (1972) describes as ‘discourses’ as discursive formations that create ‘things.’ As such, I applied FDA combined with reflexive thematic analysis with the aim to uncover organisational actors’ *stories* around the ethics of algorithmic work practice. In that regard, the FDA offered me a philosophical lens to understand the relevant discursive formations around the ethics of algorithmic practice. And concurrently, the reflexive thematic analysis provides a systematic framework on how to approach, condense, and disseminate those relevant discursive formations.

4.5.2 Analysing the Interview Data

The main method of data analysis for this research is based on FDA. But before that, the researcher has used a coding process (Miles et al., 2018) by adhering to the *reflexive thematic* technique suggested by Braun and Clarke (2019). Hence, two types of analysis were applied to analyse the findings: First, I subscribed to reflexive thematic coding to

reduce and identify the most relevant remarks. Second, I used the FDA, as an approach to better understand the dominant discourses within the data that informs the ethics of algorithmic work practices. Considering case study research design, Yin (2018) also outlines that the analysis techniques suggested by Miles et al. (2018) are not mutually exclusive, and researchers can use them in combination with other methods. Therefore, this study combines and incorporates FDA theory with reflexive thematic analysis to ensure that Foucault's key concept of *Discourse* is not ignored and can be identified through thematic analysis. The following section outlines how reflexive thematic was combined with a coding strategy to structurally implement FDA theory.

The first type of analysis is focused on Miles et al. (2018), who suggest that qualitative researchers should familiarise themselves with their data before undertaking the analysis process. Miles et al. (2018)'s framework on qualitative data analysis is highly endorsed by other scholars such as Robson and McCartan (2016) and Yin (2018) because it provides a useful initial framework, particularly informative to analyse data in case study research. For this research, the interview transcripts and documents shaped the main frame of data. These data sources were initially analysed by adopting the processes suggested by Miles et al. (2018). As such, I followed their techniques to first reduce and display the data and subsequently draw meaning from it through an FDA analytical lens. The analysis procedure also involved developing code and themes for better reduction of data (Creswell and Poth, 2016).

Also, the researcher did not totally adhere to Miles et al.'s (2018) data analysis approach due to ontological clashes between Foucault's social *constructivism* [as well as the researcher's] and the *realist* standpoint of Miles et al. (2018). The structural approach suggested by them encourages researchers to develop and utilise practical standards and techniques and leave aside others as long as high-quality conclusions are achieved. This is in contrast with the Foucauldian approach to the analysis of discourse or discursive formations. To explain, the FDA centralises on the social, political subjects or people: It problematises how and under what conditions subjects' discourses can constitute and materialise social phenomena (Ahonen et al., 2014). It means that for Foucault, all discursive statements, inscriptions, and discourse formations from subjects have weight and should be taken into consideration when one is exploring multiple 'truth(s)' of the social world (Graham, 2011).

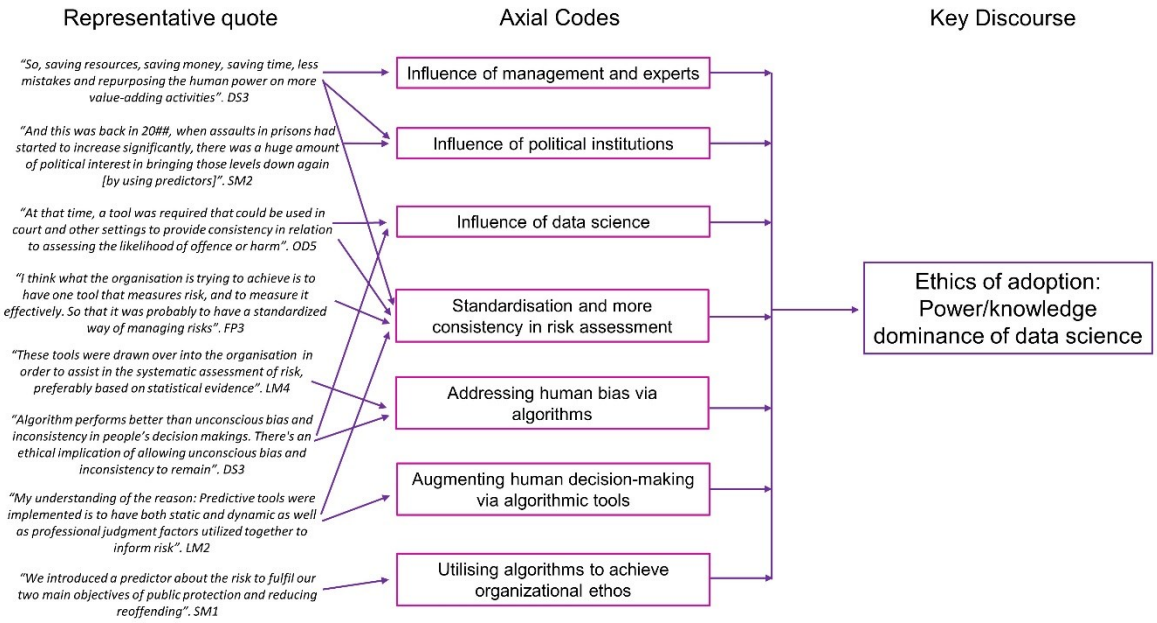
It is recommended for the researchers to code the data as soon as possible in order to elevate their understanding of the data and better theoretical building (Clark et al., 2021). Following this advice, I have aimed to code the transcripts as near as the completion of each

interview. To do so, I have applied three coding techniques suggested by Strauss and Corbin (1990) as preliminary methods, which encompass breaking down the data while aiming to categorise and conceptualise it at a later stage. The first coding technique is *open coding*. To do the open coding, I looked for the particular statements or units within the text that were relevant to the general research agenda and/or could provide novel theoretical ideas for the existing literature (Bell et al., 2018). The open coding procedure was conducted simultaneously as each transcription was completed; therefore, a more sophisticated strategy to coding was required that could explain the relationship between the discourses and theory and literature. To do so, I have applied the *axial coding* technique as the second phase of coding. In axial coding stage, I looked in-depth into how the generated concepts from the open coding phase can be correlated to other concepts, weighing them in terms of their theoretical significance. Finally, the codes that emerged from the axial coding stage are merged into a single key (core) discourse. Each key discourse identified in the findings shows the connection to the axial codes and indicates a particular explanation in relation to ethical dimensions of algorithmic work practice. These key discourses are fundamental aspects that add to the existing theory, i.e., contributing to the existing literature. It is essential to highlight that as the analysis of data is based on the FDA, these three coding techniques were used in sequence to reach a key discourse that demonstrates the organisation's or actors' perspective(s). However, for the concluding discourse, which is a crucial one, underlining the overall verdict of different stakeholders on the future landscape of the algorithmic service, I have not used the axial code approach. For this particular discourse, I have divided key findings into two contrasting discourses, depending on the frequency and repetition of sentiments. For more details, please refer to the last illustration in appendix 6.

To better conduct the coding process, I subscribed to the recommendations by Bryman (2016) and read each interview transcript and document several times. By doing this, the researcher can immerse themselves in the details of interviews and make sense of them before going deeper into them (Agar, 1996). After a few readings through the transcripts, I looked for the statements and enunciative patterns that are being repeated throughout the text. More importantly, I exercised more cognisance around the unit of analysis of this research and focused on the particular discourses that informs the ethics of algorithmic tools in each transcript or document. I also continuously reviewed the codes that have already been developed or were being developed alongside the ongoing coding process. As such, I was able to delete, merge, or redefine codes in some instances. Concurrently, I considered important Foucauldian concepts such as the existence of power dynamics within language (Fairclough, 1993) and the art of governmentality during the

process of coding. Figure 4.2 provides an example of the coding process in relation to identifying the dominant discourse of data science power/knowledge. Figure 4.2 does not represent all the codes from all 38 interviews and additional documentary sources. Moreover, figure 4.2 does not demonstrate the initial *open coding* process as this one was carried out on NVivo to gain a preliminary understanding of the repeated concepts and sentiments. Screenshots of open coding are provided in Appendix 5. For more details regarding the reflexive thematic examples, please refer to Appendix 6.

Figure 4.1 An example of reflexive coding for the first finding



To better organise the process of coding, all interview transcripts and documents were transferred to NVivo version 20. NVivo is a computer software package specifically designed to enhance and simplify the coding process and organisation of qualitative data (Bryman, 2016). Not only did the use of NVivo help me to better manage the time during the coding and its retrieval, but it also made the process more efficient and explicit. As Bell et al. (2018) argue, the use of computer software programmes can enhance transparency and provide a better audit trail on analytical processes of data. Which can then respond to the criticism on the lack of clarity of qualitative data analysis methods (Bryman and Burgess, 2002). Utilising the reflexive thematic method together with the coding technique has helped the researcher to better apply FDA theory during the analysis and draw meaningful interpretations from the data.

4.5.3 Data coding by considering the Reflexive Thematic method

I employed the coding process as a strategy to combine FDA with reflexive thematic analysis. The process of coding involves assigning codes or names to different chunks of texts, such as words, sentences, or paragraphs (Miles et al., 2018). The codes are identification titles that at best capture the meaning of the data that they were applied to (Bell et al., 2018). For example, to illustrate that the frontline practitioners' discourses were excluded from decision-making to introduce ARA tools, I have coded relevant talks under 'missing voices on algorithm introduction.' Similarly, to demonstrate the reasons for the utilisation of the ARA tool, I have gathered relevant data under the code of 'motivations for the adoption of algorithms.'

Although the coding process is mostly aligned with grounded theory data analysis (Strauss and Corbin, 1990), it is suggested that coding can be applied to any form of qualitative data analysis (Bell et al., 2018). Furthermore, coding technique is highlighted by Braun and Clarke (2019) as a reliable method in carrying out reflexive thematic analysis. There are indeed scholarly arguments that are against the codification of data when research is heavily entangled with the analysis of discourse or language (e.g., Gill, 2008). However, discourse analysis scholars such as Keller (2012) argue that although discourse analysis focuses on microstructural and semantic aspects of text (Graham, 2011), coding can help to better ascertain the causal relationships within the text. Therefore, a coding procedure was used to systematically and thoroughly approach the FDA throughout the data analysis stage of this research.

4.6 Conclusion

In this chapter, I highlighted the methodological approach used to explore the key organisational players' perspectives of algorithm ethics in workplaces. I explained how interpretivism epistemology and the use of qualitative research strategy were deemed most appropriate to address the research questions of this study.

A case study design consists of one criminal justice organisation in the global north (Europe continent) that was selected to cultivate a deeper understanding of the ethical impacts of the adoption of algorithms for risk assessment work practices. Also, this case study highlights the nuanced aspects of employee agency and resistances against algorithmic work, which can be tied to the scholarly conversations on algorithm ethics. I incorporated semi-structured interviewing as the main method to collect primary data. In addition, I gained access to a number of organisational documents to further complement the collected data from the interviews.

In order to undertake the data analysis, I have subscribed to FDA theory. As it was discussed previously, FDA is not *per se* a framework that offers a step-by-step guideline on how to analyse data, but it is a way of thinking suggested by Foucault to challenge the conventional ways of analysis and practice (Foucault, 1988). However, in order to systematically analyse the research data and draw findings from it, I combined FDA with the reflexive thematic technique. To do so, I have carried out the traditional coding processes used in many qualitative studies. The coding procedure, however, was not solely directed to dig up repetitive patterns, similarities, and contrasts within data and their relevance to the literature. But it was coordinated in way to treat data as discursive systems or rather, power formations for identification of Foucauldian notions such as governmentality, subjectification and resistance, and answering the research questions.

The following chapter, chapter 5, explains my findings on how specific power/knowledge dynamics have steered the deployment of ARA tools and how the organisational players have demonstrated novel discourses around these tools' ethical implications.

CHAPTER 5: Findings

5.1 Introduction

This chapter explains the distinctions between three competing discourses concerning the ethics of algorithmic work practices derived from the data analysis. The findings are based on the analysis of reflexive thematic analysis and Foucauldian Discourse Analysis (FDA) through a 3-stage coding procedure. They highlight three dominant and conflicting discourses that will address the research questions. Please refer to Appendix 6 for further details regarding the reflexive thematic – Foucauldian data analysis and the extracted discourses from transcripts and documentation. Appendix 6 provides some representative quotes from the interviewees as well as the documents that are used to develop the key findings of this chapter. This chapter highlights three main discourses outlined by key organisational stakeholders' relation to ethics of adoption, utilisation and work transformation via algorithmic tools.

5.2 Ethics of adoption: Power/knowledge dominance of data.

This section elaborates on the particular ethical aspects around the decisions to adopt and implement ARA tools and shows how particular rationalities advocate utilisation of ARA tools for the delivery of ethical practices.

5.2.1 Influence of management and experts

Criminal justice organisations are amongst the key components of western public administration practices. This means such systems are predominantly shaped and influenced by existing political institutions as well as the decisions made at ministerial or parliamentary level.

In that regard, findings suggest that particular discourses from political institutions initiated a change with the aim to reduce costs and transform the rehabilitation practices and interventions. *"The desire was to reduce the cost of running the service"* as highlighted by a data scientist (DS1). Furthermore, the data suggests that the organisation was keen to reduce the time spent on each case by the frontline employees. As such, another data scientist (DS4) indicated, "*We're talking about transforming rehabilitation when they [political decisions] were splitting [privatising]. I mean to say that [algorithms] save money actually"*. Another data scientist (DS8) mentioned, "*we [criminal justice] are making a lot of decisions better and a lot of decisions better because we're using data*", and the organisational *"is definitely going in the right direction in adopting more [algorithmic] risk predictors."* As finding

outlines, political decisions – which were made externally – advocated the data/evidence-based risk assessment, utilisation of intelligent algorithms, and even outsourcing [privatisation] part of the risk management practices. This was important because the influence of political decisions to partially privatise the organisation was not entirely supported by all frontline members. As a senior manager (SM3) lamented, "*The best example [for the influence of politics] – and also the worst thing that happened to our organisation probably – is when it was partly privatised years ago."* It indicates that not all members of the organisation were consulted when major decisions were made to privatise or deploy ARA tools.

Furthermore, internal investigations carried out by the service itself have identified that the amount of time being used by a practitioner to process a client's case is quite substantial. Moreover, the issue of workload coincided with a significant rise in the number of offences and the need to process offenders in swifter ways. In that regard, a senior manager (SM2) highlighted, "*We've got a rough estimate that we think that [new] ARA tool could save the frontline practitioners a lot of time in a year. And that is quite a lot of worth to us*." Therefore, ministers and other concerned political parties supported the decision to integrate risk assessment practices with algorithmic tools with the hope that intelligent technologies would streamline case allocation.

In addition, key organisational stakeholders who were involved in decision-making strategies to implement ARA tools were the people at *"senior management level,"* as indicated by a data scientist (DS1). Subsequently, those decision processes were formally approved and endorsed by policymakers and government ministers. In that respect, a senior manager (SM1) mentioned, "*The principle of introducing the tools definitely, led by senior officials within the transforming rehab programme but endorsed by ministers at the time."*

The frontline practitioners, line managers, and other professionals and even the relatively less-experienced data scientists, however, were not certain how the decisions were made, and who the key decision makers were. As a frontline practitioner (FP3) agreed, "*It **could've** been the policymakers, the civil servants, and it **may** even come directly from the government and they will put forward a motion in the [parliament], try and get votes*." Notably, some of the organisational documents indicate that the adoption of ARA tools was introduced in order to assist the frontline staff. As outlined in a policy document (OD7), "*It is designed to be an integral part of the work which practitioners do in assessing offenders; identifying the risks they pose, deciding how to minimise those risks, and how to tackle offending behaviour effectively*." However, as findings indicate, not all frontline members have been informed or briefed on the reasons for utilisation of ARA tools. Moreover, some

frontline professionals expressed doubts about the lack of their involvement in any initial decision-making processes that led to the emergence of ARA tools. With regard to decisions to introduce ARA tools, a line manager (LM5) highlighted, "*Things happen to the service because the government decides that what's gonna happen, without frankly much negotiation"*. Some frontline staff believe that the service exists to respond to the government policies which will then have a massive impact on the organisation. Therefore, the decisions have been made without much negotiation with the frontline practitioners or other professionals. Another frontline practitioner (FP5) told me, "*I'm a civil servant, and it seems that offering opinions is generally viewed as political. And practitioners' involvement in decision-making is virtually non-existent."* The top-down managerial dynamic does not allow much space for consultation with frontline practitioners. As the statements suggest, not much feedback or comments from frontline staff – who are the main end-users of ARA tools – were included in the design and introduction of these tools.

This highlights a new discourse characterised as a new dynamic of administration, control, and normalisation over the employees' working lives (Knights, 2002) and bodies (Foucault, 1977). This novel dominance is empowered by intelligent algorithmic work (Newlands, 2021). As data suggests, in the process of adoption of the ARA technologies, the policymakers were influenced [objectified] by the advantages of scientific methods and the algorithmic augmentations of the workplace. Moreover, the policymakers were seeking potential avenues to reduce the costs of running the service. Furthermore, this change disturbed some of the frontline practitioners because the criminal justice service "*plays a vital role in protecting the public from people who have offended, and we cannot hope to do this effectively without understanding the risk presented by those we manage*" as signified by a data scientist (DS3). Thus, effective risk assessment and management are clear priorities within the service strategy. On the one hand, the ethos, values, and principles of judicial systems constituted a discourse that augmentation in risk assessment was viewed as essential. And on the other, the key policymakers, senior management, and organisation officials – embraced the analytical power of algorithms – approved, supported and enacted a novel power/knowledge discourse for risk evaluation. Within this new discourse, the immediate target was the traditional format of risk assessment. The traditional form of risk assessment undertaken by frontline employees was meant to be more standardised, as the findings highlighted. In this new discourse senior management and policy makers aimed to regulate and control practitioners' work practices through three methods. Firstly, the frontline practitioners are objectified because their risk assessments were analysed and flagged up as inconsistent. Secondly, through showcasing peer-reviewed research outputs, ARA commodities intensified the power/knowledge exercise over practitioners' working lives. This

intensification was backed up by academic evidence that the traditional mode of risk assessment can sometimes be inconsistent or untransparent. Therefore, the solution was to introduce the ARA technologies to ensure a transparent, evidence-based, and ethical practice. And finally, as findings suggest, the viewpoints from frontline professionals were hardly considered as prerequisites in the ARA implementation decision-making. This asymmetrical decision-making signifies that the organisation has sought and advocated a dominant discourse of expertise of data scientists. The institutional politics and senior leadership, therefore, have heavily relied on science and relatively marginalised the inputs from frontline members.

5.2.2 The influencing discourse of data science

The data demonstrates the dominance of power/knowledge discourse (Fairclough, 1993) that is heavily influenced by the existence of data science [the actuarial data science]. This specific discourse has promised regulated offenders' case allocation, enhancing consistency and augmented decision-making for frontline practitioners in criminal justice settings. As highlighted in a relevant document (OD5), "*The ARA tool is designed to help practitioners make sound and defensible decisions.*" The data scientists and engineers per se do not possess the power to steer the adoption and implementation of ARA technologies. However, through research-based evidence, a novel power/knowledge has formed that signified frontline staff's inconsistent assessment and promised enhanced offender management via algorithmic tools. To embed this inconsistency, another policy document mentions (OD6), "*a tool was required that could be used in court and other settings to provide consistency in relation to assessing the likelihood of further harmful offending.*" It highlights that the organisation knew and emphasised that there was a need for a system to detect the "*most dangerous offender,"* as a data scientist (DS5) said. Furthermore, the research conducted in-house by data teams has highlighted some issues of inconsistency in the risk assessment amongst the frontline officers. In that regard, a data scientist (DS1) mentioned, "*We knew from previous work of mine that the professional judgments - the judgments that practitioners make about 'who is the most dangerous' - can be* ***inconsistent.****"* Hence, the decision was made to design an algorithmic tool that can help the frontline professional at the outset of their risk assessment. Thus, according to the power/knowledge of data science, an ethical service can be achieved via the utilisation of algorithmic tools.

The findings illuminate that *data* plays a crucial 'discursive' role in the risk assessment processes*. "We believe that* ***Data,*** *in the very broader sense, can help and support of those professional judgments*", highlights a data scientist (DS3). The data

scientists have been the pioneers to incorporate the data and use it for building the predictive algorithms. Characterised as experts, a strong discourse exists amongst them in the sense that integrating descriptive statistics with algorithms, which they believe will potentially help the frontline officers in better assessing the risk.

These statements indicate that although the motivation to employ ARA tools was catalysed by the availability of data, there was also the influence of scientists and their power/knowledge expertise that paved the way for ARA. Their dominant discourse has highlighted ubiquitous inconsistencies within human practitioners' judgements, denouncing the traditional format of human-based risk assessment. In line with this change, a few of the frontline practitioners (and line managers) were objectified by the highlighted similar ideas. As the analysis suggests, many frontline practitioners understand and acknowledge that consistency and standardisation of risk assessment were, indeed, amongst the main reasons for the adoption of algorithmic tools. Integration of data, statistical reasoning and algorithmic augmentation have assisted frontline staff in their professional judgment. For instance, a line manager (LM7) indicated, "*I think it [the ARA tool] helped in terms of getting some **consistency** in our approach to risk assessment and the likelihood of reoffending primarily."* Some frontline practitioners also underscored that the criminal justice organisation is probably keen "*to have a **standardized** way of managing risks",* as mentioned by a frontline practitioner (FP4). A line manager (LM1) also mentioned that "*incorporating data to measure risk"* shaped the ARA tools in order to have a universal, harmonious method to predict risk. In that regard, according to the findings, ethical practice for some frontline employees is crystallised via the use of algorithmic tools, as the tools provide them with less disproportionate predictions.

It was reported by a data scientist (DS1) that the utilisation of ARA technologies in this criminal justice organisation "*has never been with the intention to make frontline practitioners redundant.*" Algorithmic tools were never designed to totally automate professional judgement and replace employees. On the contrary, they were designed to act as guidance or a starting point for the professional judgement of practitioners or to provide them with a statistical baseline and a reliable assistant at the outset of decision-making. Hence, the dominant data science discourse was introduced to make the risk assessment more efficient, or rather, provide a basis for defensible decision-making. Ultimately, the findings indicate that scientific power/knowledge became the dominant discourse via instruments including academic research outputs, efficient utilisation of data, and in-house engineering capabilities. Thus, the use of algorithms was considered an appropriate method to make risk assessment practices more ethical, consistent, and cost efficient. The influential discourse of data science in this organisation is ubiquitous and has produced an

unprecedented discursive power/knowledge dominance (Raffnsøe et al., 2019) that advocates human-algorithm partnership as a solution to overcome ethical issues.

5.2.3 Addressing human bias via algorithms

Human bias is viewed as an ethical challenge and a justification for using algorithmic tools (Charlwood and Guenole, 2022). In that regard, different organisational stakeholders have considered human errors and biases as serious ethical issues that need to be addressed. The senior management is predominantly inclined to utilise algorithms to minimise the risk of unconscious bias. As a senior manager (SM2) explained, *"[The organisation] is aware that professional judgement can be disproportionate to certain groups of people.*" This viewpoint was also reciprocated by the data scientists arguing that algorithms are able to rectify the issue of bias. A data scientist (DS9) mentioned, "*The other side of the coin is that people are ignoring the algorithm when actually the algorithm performs better than staff given their unconscious bias in decision making.*" Some frontline practitioners also expressed that human-based risk assessments can sometimes be inflated by bias or stereotypes. For example, a frontline practitioner (FP7) mentioned, "*The practitioners like to think they are non-biased, but everybody knows they have their own biases.*" Or another frontline practitioner (FP8), who also gatekeeps the cases, explained, "*People within the government thought that we [practitioners] were biased. There was unconscious bias, which I think is right. There is now. Human beings' professional judgements are not totally objective. They can be subjective, where actuarial data is fact.*" With regards to unconscious biases, another frontline practitioner (FP12) also observed, "S*ome assessments were not necessarily reflective of that individual (the offender); there's obviously been a rupture in the [offender's] relationship, and it's shown in the assessment and had resulted in a biased judgment.*"

According to the findings, dominant discourse of data science seems to have formed a belief amongst some frontline practitioners in the sense that their professional risk evaluations can be tenuous including potential biases and/or prejudices. The findings suggest that the new dominant discourse insinuates this notion that practitioners' assessments might not take into consideration all the necessary factors needed to make objective assessments. And this dominant power/knowledge is backed up by scientific methods and supported by the senior management. Subsequently, this discourse has made some frontline staff believe that their work might be skewed and inconsistent.

As data indicates, many of the frontline end-users of ARA have acknowledged the benefits of algorithms. They have explained how ARA tools endowed them with more impartial/objective predictions than theirs. Many practitioners feel more confident having

ARA tools assisting them with in their decision-makings. Hence, it is ethically necessary to have a platform that can provide them with a holistic view of each case. As a practitioner (FP9) argued, "*It [the ARA tool] helps you feel safer in your assessment because you can worry sometimes that you're not wholly objective."*

My findings indicate that human bias is another contributing factor for the organisation to integrate ARA tools in risk assessment practices. The frontline staff have predominantly acknowledged this shortcoming in their work and are inclined to rely more on the outputs of algorithms. Utilising algorithmic predictions gives them more confidence and provides them with scientific evidence to defend their assessment. As a policy document illustrates (OD2), "*Using the [data-driven tools] could help you [practitioners] understand the impact of your intervention and could also help to demonstrate this to others*."

Hence, dominant scientific power/knowledge possesses the power to dominate work practices which have unleashed it upon the frontline people. It is, however, not accurate to link power dominance to senior management, ARA designers, or data scientists. This is exactly the pitfall Foucault has warned us about (Knights, 2002): Power is not a 'thing' that can be possessed, owned or transferred by individuals, as opposed to the concept of sovereign power (Foucault, 2020; Barratt, 2008). It is a cycle, or rather a dynamic, that is being exercised over human bodies; it regulates them and constitutes them as objects to expand the domain of knowledge (Heller, 1996). Thus, examining power relations through algorithmic work practices, one needs to account that these tools are not solely designed to streamline the risk assessment process and bring more consistency into the system. They also function as instruments in order to make practitioners self-disciplined and self-regulated (Introna, 2016). As data outlines, in this organisational setting, many practitioners have willingly welcomed and utilised algorithms because they want to deliver an ethical, bias-free risk assessment.

5.2.4 Utilising algorithms to achieve organisational ethos

According to data, another important reason that has catalysed the development of ARA is to protect the society from harm. This discourse was much more conspicuous amongst the organisational senior leadership, data scientists, and some of the line managers. Protecting the community from harm has been amongst the main tenets of many western criminal justice systems and has been used as a pretext to further justify the implementation of ARA tools. As a senior manager (SM1) agreed, *'Our two main objectives are **public protection** and **reducing reoffending.**'* Thus, utilising ARA tools has been considered as an ethical solution to fulfil the organisation promise: *'Prevention of reoffending, protecting of the public'*

as explained by another senior manager (SM2). Findings suggest, the organisation agreed on the introduction of data-driven algorithmic tools since the tools can enable the service "*to ensure that decision-making is based on sound insight.*" By doing so, ARA tools can "*strengthen the wider justice system, helping us [the organisation] to deliver on its goals: reducing reoffending, providing swift access to justice, and protecting the public*" as outlined in a policy document (OD10).

According to the findings, some of the frontline practitioners have also reiterated a similar sentiment. As such, a frontline practitioner (FP8) explained, "*The management of risk for the public protection work and rehabilitation of offenders sometimes can go hand in hand."* Protecting the public has significantly transformed through the use of ARA tools. Some frontline practitioners praised the utilisation of ARA tools, suggesting that they yield more fruitful outcomes with regards to protecting the community. This criminal justice organisation, thereby, has managed to foster the acceptance level of ARA tools amongst the practitioners since it has envisaged the fulfilment of organisational ethos through algorithms. In other words, algorithms were seen to offer solutions to ensure that not only the risk assessments are ethical, but also to "*enable access to higher quality and more appropriate rehabilitative interventions*" according to a strategy document (OD4).

5.3 A discursive shift amongst some organisational actors

In this section, I highlight the opposing influential discourse showing the growth of a steady discursive shift amongst organisational actors in relation to their work interaction with ARA tools. This discursive shift, as the findings suggest, is developed through organisational actors' awareness and reflexivity towards ARA practices. The following sections bring to light the concerns in this dominant discourse.

5.3.1 Concerns on utilisation of ARA

As the findings indicate, the implementation of ARA tools was received with mixed views by the practitioner end-users. Whilst the introduction of algorithmic predictors was mostly embraced by many groups of frontline practitioners, there were those who did not immediately trust and accept the idea of ARA practices: As a senior manager (SM1) explained, "*I remember very vividly doing a demo for a group of line managers of the new [ARA] tool and just seeing the expression on their face as they just were delighted, and they were thinking this [ARA tool] saves them so much time."* Some practitioners who had worked in the service for years were somewhat sceptical of ARA technology and hesitant to accept it in their jobs. For example, the mentioned senior manager (SM1) referred to

practitioners' scepticism and questioned, "*How can a computer understand whether a real human being is a risk of hurting others."*

Algorithmic tools, according to some organisational actors, are deemed as double-edged swords: On the one hand, they have brought many attractive features to traditional human-based risk assessment, including bias-free or swift case processing. As such, due to streamlined risk assessment, many practitioners have complied with the new ARA modalities and placed their trust (Robinson, 2020) in these tools. In that regard, a data scientist (DS5) agreed, "*The feedback has been very successful. What they [practitioners] say is two things: One, is that it [the ARA tool] is taking away the cumbersome and time-consuming process of doing manual job. And two, particularly in terms of the algorithm, is that it has helped us with gaining a better vision of our job [of risk assessment]*."

On the other hand, the findings show some indications of concern around the utilisation and practicality of ARA tools. The emergence of ARA tools was received with cynicism in the sense that computer intelligence might not be able to capture all risk factors that trigger an offence. In other words, a number of frontline professionals, at the outset of introducing algorithms, were able to identify and highlight the ambiguities in the tools. Subsequently, some stakeholders took action and expressed their concerns on ARA tools. For instance, a frontline practitioner (FP1) mentioned, "*[ARA tools] do not take everything [various psychological factors] into account.*" Indeed, the frontline practitioners are generally encouraged – or rather required – to use predictions from the ARA tools. But some of them have their doubts. In line with this, another practitioner (FP11) expressed, "*Human professional judgement is better at predicting risks*". These practitioners believe that human professional [clinical] judgement is still needed. A line manager (LM5) agreed, *"[Human beings] are not apples. There are dynamic things that change day-to-day, week to week. And sometimes actuarial [ARA] tools can't pick it up.*" In that respect, an internal survey document (OD4) reiterates practitioners' doubts, "*how much professional [human-based] assessment is required and whether more should be used via algorithms.*" Such sentiments show that, in some cases, the prediction of ARA has caused confusion amongst the practitioners.

The findings illustrate that a few practitioners initially saw the emergence of ARA systems as a hindrance to their professional judgement and a threat to their agency. At the outset of the algorithmic transformation, the style of contest against technology was much more collective [and tangible]. For example, the raised concerns were conveyed through trade unions. A senior manager (SM2) highlighted the early contestations, "*The concern expressed by frontline staff, and therefore their trade unions, was that this [adoption of ARA*

tools] was basically dumbing down the role of staff." In other words, some frontline staff were concerned that their power/knowledge and working experiences may be replaced by algorithms. Similarly, a line manager (LM5) mentioned, "*Trade unions were very concerned about our professional judgement, professionalism, and social work. And when we got these actuarial [ARA] tools, it had a bit of clinical [human] assessment.*" It suggests that a number of frontline people were concerned about how their future landscape of risk assessment will be as algorithms were transforming processes.

Issues such as algorithm trustworthiness (Leicht-Deobald et al., 2019) and aversion (Dietvorst et al., 2015) are ethical dilemmas that may affect our interactions with algorithm agents (Burton et al., 2020). Findings also indicate many frontline staff felt that ARA may undermine their professional judgement. Moreover, findings also illustrate particular concerns in the sense that practitioner end-users felt a potential danger to their professionalism and were keen to voice and discuss this issue (Busch et al., 2024) in the beginning of utilising algorithms. The findings also identify that some practitioners opposed the use of algorithmic technologies via contesting or averting them.

5.3.2 Concerns on lack of interdepartmental collaborations

My analysis explains that communication, collaboration and interactions between teams and departments are not optimal, particularly in relation to ARA technologies. For instance, there is an unsubstantiated claim, especially amongst some data scientists, that the majority of frontline staff do not understand and appreciate ARA tools. As a data scientist (DS5) claimed, "*People [practitioners] are reluctant to support and welcome technology; I've noticed that they think: Oh, data team is doing what we already been doing; they trying to prove that they can do better than us.*" Notably, the data science team believes the reason ARA systems do not resonate well with the frontline staff is because of their unwillingness to embrace and trust the intelligent technologies. Another data scientist (DS1) agreed, "*ARA tool is clearly not working,*" which means some frontline people do not trust it. As such, the data science team argues that many frontline members do not understand how/what the ARA tools actually work or predict. A data scientist (DS1) mentioned this sentiment: "*They [frontline practitioners] haven't read the risk assessment guidance, which has been out there for two years, which tells them to start with the algorithmic score, add in your observations, end up with the risk assessment.*" And this lack of trust, according to the data team is blamed on practitioners' inadequate engagement with ARA tools and their user manuals. In that regard, a data scientist (DS7) highlighted, "*[Practitioners] really don't know what the algorithmic output actually represents."* And this issue has led to further aversion, as another

data expert (DS8) mentioned, "*They [practitioners] assume that actuarial methods will be far inferior to their professional judgments."*

That being said, the data science team admits that frontline employees' lack of trust in algorithms is linked with ineffective communications between organisation teams. A data scientist (DS4) highlighted that "*we work on the feedback provided by the frontline staff, and they work to make the algorithm better, as they think it is within their ability to change and improve."* It means that the data team is keen to better communicate with frontline people and believes that their feedback should be taken into account.

The findings also encompass the remarks from the frontline practitioners on the lack of intradepartmental collaboration. As analysis suggests, employees have viewpoints and are keen to voice them to senior management and data teams. However, there are ethical issues that have affected the communication between these key stakeholders. A frontline professional (FP2) explained, "*There isn't any collaboration between us and data teams…That doesn't exist."* According to the data, some frontline members indicated their uncertainty about the existence of a data science division in their organisation. Indeed, the staff members are aware that the design/development of algorithmic tools is based on scientific research. But they were not entirely certain of the extent to which the justice system has been involved in the design process and who the key designers are. For instance, a practitioner (FP15) mentioned, "*I don't think I've had any personal direct involvement with people that are involved in the [ARA] predictors, and I'm not sure if that's because I joined the service after you know all the [algorithmic] changes.*" Some of the frontline employees who have been in the service for years did remember how computerisation and digitisation had taken place. As they recall, like any other organisation, the criminal justice system has gone through numerous transformations, from using physical copies of reports and multiple files and folders to the emergence of computerised management systems and the recent changes due to algorithmic risk predictors. However, not many of them had had any direct or indirect interactions with the data science team. "*I don't think there's an awful lot of joint interaction all the time,"* explained by a frontline practitioner (FP3). Similarly, another practitioner (FP11) highlighted, "*I mean, there's not a lot of direct interaction between the actual data scientists and ourselves."* The findings indicate that although the ARA technologies are developed in-house by criminal justice service itself, many practitioner end-users are uncertain how they can interact or engage further with data division team. Effective collaboration between parties can contribute positively ensuring the utilisation of ethical, user-friendly algorithms (Chowdhury et al., 2023).

Consequently, as the findings demonstrate, many frontline employees have been unable to engage effectively with ARA tools because they are unable to get what they want from it. As explained in the quotes, many practitioners are unaware of the avenues to get in touch with data science team and convey their feedback easily. This can be considered another aspect of algorithmic power/knowledge dominance that objectifies and circumvents individuals within its domain (Introna, 2016). In this context the practitioners (end-users) are constituted as bodies to expand the knowledge of algorithmic orthodoxy whilst their working lives are being controlled by the dominant power (Mennicken and Miller, 2014). Frontline members who participated in the study almost unanimously mentioned that they cannot recall any communications with the data analysts. If they face any difficulties in their interactions with ARA tools, they can raise that with their line manager. A frontline practitioner (FP10) agreed, "*I just will tell my manager, and whether it conveys or gets conveyed or not, I don't actually know, because I've never really seen things change.*" The organisation has required practitioners to utilise the ARA scores as the baseline whilst it is difficult for practitioners to add their own perspectives in the ARA tools. The findings indicate that these frontline members are asked why they need to use ARA, but their feedback may not change anything. In that regard, a frontline practitioner (FP13) lamented, "*I don't think much change will happen even if we do complain."* The organisation rarely listens to the viewpoints from lower levels of the organisation including the frontline employees. When it comes to a substantial change that would affect the working lives of employees, there is not much pre-examination or negotiation. It is more about *"responding to government policy so it can have a quite massive impact on practice"*, as a senior manager (SM3) mentioned. One HR manager (HRM2) referred to the utilisation of ARA tools, mentioning, "*Things happen to our service because the government decides that that's what's gonna happen."* Another line manager (LM6) blamed the lack of time and volume of caseload, explaining, "*we particularly don't have the time to raise things about the tools*". It indicates that algorithmic transformations have had significant impacts on this organisation. Nevertheless, many organisational actors have not been able to express their views around these changes, or rather, have never been asked.

As findings indicate, many frontline practitioners have sought to communicate with the designers of ARA tools, i.e., the data science team, yet their efforts have not been very fruitful. Indeed, the designers have been able to provide practitioners with relatively basic material on what ARA is and how to use it. Yet, this also has been limited, and not much interaction has taken place since the ARA tools were deployed. In contrast, findings also highlight, much feedback and constructive comments exist from practitioner end-users that can help to foster the effectiveness of ARA tools. The feedback from these practitioner end-

users may even help to tackle any potential ethical dilemmas of algorithms as discussed by Ananny and Crawford (2018). As suggested by Lepri et al. (2018), ethical issues such as in-built biases within algorithms can reproduce and further expand due to algorithmic decision-making. Thus, those who work with algorithms, which in this case, are the practitioners can detect these issues and transferred them to data teams. Yet, as findings point out, due to limited engagement between these stakeholders, many voices remain unheard.

Overall, many practitioners have shifted from the dominant discourse and have been scrutinising and criticising ARA technologies. These criticisms are neither instances of algorithm aversion, nor a total resistance. But it is the growth of awareness from a total algorithm appreciation to a point where actors have been able to unearth the shortcomings or potential flaws in ARA tools. As such, the practitioners are no longer the mere end-users of algorithms circumvented by algorithmic agential power (Introna, 2016) but are subjects who can exercise their agency to foster their interactions with these tools and engage in an ethical practice (Skinner, 2013). In the next and final section, I elaborate further on the discourses of organisational actors' awareness and agency and explain the nuances of ethics identified by these actors.

5.4 Ethical nuances within discourses of key organisational stakeholders

In the final section of findings, I explain the main points raised by key employees of the organisation that show the emergence of a discursive shift. These points highlight how the actors, specifically practitioner end-users, have become aware of and challenged the dominant discourse. This section reveals particular nuances in the discourses of these actors, from identification of in-built ethical issues of ARA tools, to ethics of their work practice, to understanding the impacts of the tools on themselves and their professional work practices. This section demonstrates employees' awareness, self-reflexivity, and subjectivity through quotes and statements in order to understand ethics and address the research questions.

5.4.1 Lack of adequate/appropriate data

The process of design and development of ARA tools was due to the availability of data that existed in the criminal justice organisation. In other words, the idea was to utilise the existing data from the cases, police records, and other relevant sources to develop a tool capable of analysing potential risks or rather "*predicting a future event",* as mentioned by a senior manager (SM1). However, as the findings outline, some employees raised doubts around this idea. In that regard, the issue of data limitations was put forward by some data

scientists. For instance, a data expert (DS2) said, "*Sometimes we even have duplicate records for the same person with different parameters, which is dangerous when computers are consuming that and making calculations based on that.*" A user guideline (OD5) about ARA tools seems to point out the same issue, specifying that "*the ARA tool predicts proven serious reoffending, with some limited exceptions. No actuarial risk tool can predict undetected reoffending.*" According to this document, the organisation agrees that there are flaws with the ARA systems. Thus, algorithmic intelligence at the moment is not fully capable of predicting an act of violence. However, as the previous sections indicated, the organisation argues ARA tools are effective solutions for predicting risks, despite the lack of accurate algorithmic predictors.

As findings suggest, data scientists also acknowledge the limitations within the data used to develop algorithms. A data scientist (DS2) agreed, "*So you've got lots of patchy data. That's as good as it is."* It is a challenge in this organisational setting that a lot of data is collected for the purpures of administrating offenders and other clients, and not particularly for the purposes of creating an algorithm. Hence, the data used to make algorithms is not ideal. Furthermore, some data scientists lament the old non-algorithmic methods. As such, a scientist (DS1) explained, "*Previous methods were poor predictors, whereas the [ARA] actuarial scores are good predictors. One could form an argument that one shouldn't use the professional judgment at all."* Such statements underline that although data science team acknowledges the limitations with the data affecting prediction efficiency of the tools, some still believe that ARA project have helped the organisation. In that regard, a policy document (OD8) says, "*People want to see transparency about the data input and how this leads to the models' outputs*. *A key benefit is understood to be government employees saving time"*. It is generally accepted amongst some scientists that the existence of ARA [with all its data flaws and shortcomings] is better than a 'no technology' situation and complete reliance on human judgments calls. Nevertheless, according to the findings, it is crucial for the organisation to incorporate accurate, transparent data when developing algorithmic predictors. A policy document (OD10) highlights that the use of transparent data "*would help to address concerns about accuracy and reliability, as well as privacy."*

The people at the frontline including practitioners and their line managers, however, provide a more detailed picture of the data limitation issue. In particular, some practitioners found that there are many nuances linked with the prediction of risk which can never be captured by algorithms. For instance, a line manager (LM7) explained, "*If somebody's been a victim of domestic abuse, traumatic childhood abuses*, *relationship status, and/or even their vulnerability for manipulation, it can result in an offence, which may not be included in ARA."* Hence, the generated prediction might not offer a very accurate estimate for the

likelihood of an offence occurrence. A number of practitioners believe that there are many intangible variables that affect individuals and steer them into committing an offence. These practitioners have analysed that the generated predictions of algorithms only take into account generic demographic data. A frontline practitioner (FP4) illustrated this issue, "I *think it [data] is biased and for the simple reason. Because of research that they've used may be to do with white males. So then when we're using those actuarial [ARA] scores, an outside of white males may receive harsher rehabilitative programmes*".

Furthermore, some frontline members criticised the research that has produced the data and ARA tools. In that respect, a line manager (LM5) argued, "*We're working on research that doesn't actually take into account the advances and changes that we have to society and behaviours and socialization as a whole.*" Also, a line manager (LM1) reflected on the lack of *"contextualised societal factors in the data as something that it doesn't seem to be regularly looked at.*" Thus, whilst many of the participants, including data scientists and frontline practitioners, highlighted limitations within the data, practitioner end-users further labelled this as an important ethical issue. This is in line with the revelations of Tsamados et al. (2022) who argue that the issue of misguided or inconclusive data might question the neutrality of an algorithm. As findings indicate, this issue has become an important concern for some actors working with ARA tools and they are keen to have algorithms built upon more inclusive robust datasets.

5.4.2 Potential risks due to unfair ARA outcomes

It was argued in chapter 2 that the design and development process of nearly all algorithms and AI technologies are based on importing massive data clusters – or big data – into a computer programme (Etzioni and Etzioni, 2017). This computer programme subsequently becomes a predictive decision-making tool that mainly relies on what is fed back to it. An algorithm feeds on data that could potentially be inflated by biases, stereotypes, and prejudices. This ethical issue surfaced in this study's findings.

As discussed earlier in this chapter, the organisation has access to many datasets, including former offenders' cases, police records, and community service history, that were used to create ARA systems. As such, a data scientist (DS9) explained, "*The challenge is replicating biased decisions when you don't mean to.*" Another data scientist (DS8) agreed, "*an algorithm might be different, like having a bias towards some ethnicity and others. And if you don't have that data captured, you can't assess what the impacts on different groups are. So, I think there's a question about fairness".* As these quotes suggest, some members of data science team have understood the issue of unwanted bias and discriminatory

outcomes in relation to ARA tools. As it seems, these scientists are aware that the implemented ARA solutions are not entirely innocuous. But there are obstacles that make it difficult to fully resolve the issue, such as discriminatory outcomes. In that regard, a data scientist (DS5) explained, *"The issues around ethical implications and algorithmic biases are still relatively new at organisational level."* It means that although the premise of ethical algorithms is institutionally well-established, the organisation also needs to better understand this problem. Another data scientist (DS7) also highlighted this issue, "*this often means you need good **senior buy-in** and **support** to deliver this message [algorithm ethics] across the organisation".* Data suggests that some scientists expressed their vexation at how rudimentary the data ethics policies are and called for more action from senior management. These stakeholders argue that ethical principles around data are not yet considered as a regulatory or mandatory policy.

Some data scientists explained the measures that are taken [or need to be taken] in order to minimise the risk of algorithm bias. Firstly, they emphasised the importance of tests and trials in order to ensure the ARA tools are functioning at a desirable level. The data team believes that it is imperative to compare the predictions of an algorithm against humans. A data scientist (DS3) explained, "*One of the things that we're trying out is where you've designed something that you think can predict something, and you compare the output against when it's not being used.*" As findings illuminate, such comparisons may be effective in measuring the performance of an algorithm, but also an efficient method to unearth the ethical flaws, such as data biases. A data expert (DS8) pointed out, "*Through tests, better data monitoring can be archived which means that technical experts would be able to detect anomalies and imbalances in the data better.*" Another expert (DS7) reiterated, "*randomised control trials for some areas for the [algorithm] tools we're building to see what the impact is*". A scientist (DS6) highlights the benefits of such control trials: *"We see that [the ARA tool] is having a positive or negative impact. And whether it is actually having the desired outcomes that they want."* In line with tests around ARA utilisation, some data scientists have mentioned the concept of *algorithm sustainability* and how vital it is for having a neutral algorithm. A data scientist (DS3) clarified, "*If you just do [the design and implementation] once and then nothing happens, that is not a very sustainable algorithm."* Some scientists have subscribed to this concept, arguing that an algorithm should be in the constant process of evolution. If a predictive algorithm is expected to continue to be fair, it must be regularly updated.

Secondly, the data team indicates that other stakeholders should be trained and be aware of data issues. A data expert (DS7) highlighted, "*it's important that the concepts, principles, etc., are well socialised and understood to ensure consistency. If staff feel*

confident and empowered, they'll also be more likely to engage with an ethical framework and challenge/question where appropriate." It means that some scientists believe that having an ethical algorithmic system requires the input of all stakeholders. These data scientists seem to value other stakeholders' awareness around algorithm ethics.

The findings illustrate that the scientific power/knowledge relation has expanded its domain (Hardy and Thomas, 2014) and situates data science as its primary actor to further expand. The issue of bias in ARA has been identified, yet solutions seem to be mainly hypothetical, driven by data scientists. Put differently, one can argue that although bias has been identified and acknowledged as an ethical concern, little has been done pragmatically to resolve it. This may suggest that algorithms serve the organisational needs, regardless of their flaws. As the findings will show in the following paragraphs, some frontline staff and their line managers have uncovered and identified many nuances around the issue biased assessment. Employee viewpoints and concepts influence different discourses that can provide novel understandings around the ethics of ARA tools. Yet, those ethical discourses have been categorically disregarded and marginalised. Although the viewpoints from the data science team seem to encompass more structural – or rather scientific – aspects of data biases, they seem to only provide a holistic picture of the issue. As findings indicated, scientific ethical viewpoints are important, yet yield little on how the biased risk assessment challenges can be resolved.

Statements from the line managers and practitioners signify the other side of the narrative, offering interesting nuances around ARA ethics. Some frontline members have touched upon a more fundamental ethical issue which is best described as *feeding more bias* into ARA systems. These practitioners explained that the input data used to develop ARA tools might have been contaminated by several biases or stereotypes, stemming from the frontline members themselves. In other words, biased risk assessment will lead to the production of biased data, which is then fed back to the ARA tools, creating a feedback loop that amplifies biased assessment. A practitioner (FP3) reflected, "*It [data] could be discriminatory as well in many ways…because we're naturally complex individuals and we can have our own biases, which then affect the actual input and the data that we input into our risk assessments."* These practitioners highlight the importance of being aware of their own biases and the ways in which an assessment can be affected by biased thoughts. As findings uncover, some of the practitioners have brought their own knowledge and discourses of ethics into the risk assessments. For instance, a practitioner (FP4) emphasised, "*We [practitioners] have to reflect on ourselves and actually have an appreciation of what our personal beliefs are, and so far, as we don't contaminate the data or contaminate our own assessment.*" Similarly, another practitioner (FP6) said, "*The clinical*

assessment [or human professional judgments] may be based on stereotypical views, and our disproportionate minds can be represented in our judgment calls." Likewise, a practitioner (FP8) outlined, *"[ARA] predictions will be influenced by the elements of bias because they still depend on what's imported into the algorithms*". These quotes suggest that practitioner end-users of ARA have interesting and momentous inputs with regards to the issue of bias and overall ethical practice. A practitioner (FP5) mentioned, "*There are many reasons why the service is labelled as institutionally racist."* As findings show, some perspectives are excluded in relation to the issue of algorithm bias, and this may damage the reputation of the organisation.

The findings also point out ethical issues such as intersectionality and gender prejudices as well as disproportionate predictions due to age. A frontline member (FP11) mentioned, "*They [ARA tools] were created primarily surrounding males. So female offenders aren't taken into consideration.*" This again highlights that data used in ARA might be imbalanced, unfair or biased. A line manager (LM4) shared a similar sentiment: "*Algorithm is underrepresenting all demographics and variables of society including females.*" As an outsider to this organisation, I found treating offenders of different genders quite inconsistent: I had a follow-up question: Why should female offenders be assessed differently from males if they have all committed the same offence, and given that professional judgement is supposed to be impartial and objective? A point was raised by a frontline member (FP10) with regards to psychological differences between two genders that are not necessarily captured by ARA tools. *"We need to have something [an algorithm] that's for them because male and female and what they go through. I'm not saying like we should treat them any differently because obviously it's exactly the same. But how/what they've been through is quite different for females. I'm not saying men don't have anxiety and depression, but females have a lot of factors and a lot of external factors that have impacted their lives in so many different ways".* Some practitioners argue that algorithms are probably increasing the risk for female offenders. A practitioner (FP4) reflected on this, "*criminal justice system wasn't made for females. It was there for men because they looked at women didn't commit offences. So, when a woman comes before the court, she has stepped outside of what her gender is*". And due to such inconsistencies within data and ARA, the effectiveness of algorithmic predictors has been questioned.

A similar issue was raised around the age of those who have committed an offence. According to some observations, the ARA tools might be skewed and essentially label young offenders as *high risk*. A frontline participant (FP2) explained, "*I have people [offender cases] in their 50s and 60s…this is the second offence that it's very low. They're [the ARA output] not suitable because the score is not high enough.*" This is another ethical matter,

according to some practitioners, as those of different ages but charged with the same offences might be evaluated disproportionately. This highlights the importance of ethics because the ARA tools are utilised to maximise impartial, bias-free decision-makings. Findings show how some end-users indicate otherwise, underlining relatively inconsistent predictions. A frontline practitioner (FP1) outlined, "*ARA take short period of criminal history and because it's over such a long space of time, I find it's not very accurate*". According to some practitioners, ARA tool may assess younger individuals as higher risk because it relies on available data of the recent offences. The tool may be ignoring the psychological nature and the violence history of predators who might not have committed any offence for decades and have been recently charged with a serious one. This type of bias is yet another ethical dimension of ARA tools that is uncovered by the working experiences of frontline practitioners and is highly emphasised. Put differently, the findings suggest that working experiences with ARA tools have helped some practitioners to uncover some ethical issues that might be unknown to data scientists. The practitioners' meticulousness and unique power/knowledge are invaluable to further understanding the ethical issues that may arise from ARA practices.

As the findings indicate, it has been a difficult challenge for the organisation to try and mitigate the issues of gender/age biases, in spite of using ARA commodities. As I explained, there is a dominance of science power/knowledge that steered the processes of ARA design and adoption. A user guideline document (OD5) argues, "*ARA has been constructed to include gender as a risk factor and has been validated for women who have been convicted.*" The data science team – as the sole algorithm designers – seems to be entrenched with the existing biases of ARA and data, though succinctly admitting that the situation is not ideal. A data expert (DS1) agreed, "*We have concerns about upstream bias in the criminal justice system, we can't get rid of that entirely because people come to us with a criminal record. and we have to use that in [algorithmic] prediction".* Thus, whilst the organisation and data science team are keen to overcome bias in decision-makings, this ethical issue persists. And since the organisation is driven by scientific expertise, the utilisation of data, statistical evidence, and algorithms is considered the most appropriate solution to resolve the issue. In that respect, a policy document (OD2) highlights, *"We need to make the best possible use of our data to ensure that decision-making is based on sound insight…. Our digital, data and analysis teams will work in partnership to transform the use of data across the whole justice system."*

The findings also demonstrate a scientific, evidence-based approach has been the dominant approach to overcome bias. Furthermore, the emergence of algorithmic tools has enabled the criminal justice system to constitute standardisation, simplification, and

stabilisation, transforming human-based risk assessment. Yet, as findings show, there are interesting, underexplored viewpoints around bias and prejudice brought to light by frontline practitioners. However, these discourses are less likely to be conveyed to data experts and are predominantly marginalised. Many practitioner end-users of ARA tools have grasped, highlighted, and lamented many aspects of the algorithms. Yet, the senior management and data division's discourses advocate the supremacy of algorithms to resolve bias and other ethical issues.

5.4.3 Labelling and Categorising Individuals via Algorithms

What was uncovered in this study's findings as ethical issues of algorithms has already been highlighted in the existing algorithm/AI ethics literature. The challenges, such as biased data, gender/age discrimination, and prejudicial [or stereotypical] outputs, are highlighted in the literature myriad times (e.g., Leslie, 2019; Mittelstadt et al., 2016; Vassilopoulou et al., 2022; Dignum, 2018). However, a disturbing ethical matter was highlighted by the frontline practitioners that, to date, has not been deeply discussed in the relevant literature. As such, some practitioners explain that the utilisation of ARA tools might become a modality to benchmark or categorise individuals. On this point, a frontline practitioner (FP11) explained, "*People are feeling like a number rather than an individual [to us]. I think that was a concern that an individual was just a number up against other people."* There is a risk that ARA tools might be numericizing individuals or grouping people together. As explained previously, the ARA technologies only rely on the historical data and make predictions based on similar cases. To put this differently, the ARA tools might represent an individual as a 'number' that needs calculating.

On this point, another practitioner (FP1) mentioned, *"[ARA] reduces human behaviour to numbers… It sees them as a number. It reduces people down into their simplest forms. The people who have been traumatized; people who have traumatized other people; people who need to be seen as an individual, as a person"*. According to some practitioners, a prediction from an ARA tool provides them with a very holistic perspective of the individual. Since algorithms are not able to take into account the nuances or contextual factors, their output may be too rigid. Practitioners have raised this issue not just for the ARA tools they are using but also as a drawback around artificial intelligence technologies. Whilst some frontline practitioners are aware of this ethical issue, the data science division has almost been unaware or silent. Notably, this issue was not raised by any of the eight interviewed scientists.

Many frontline staff have realised that although they are not in a position to make changes, they still can voice their own concerns and highlighted ethical challenges that are harboured by ARA tools. They have been able to unfold this nature of ARA tools by highlighting that ARA tools can never show compassion and human connection; the ARA tools are only statistics. In that regard, the human professionals exercised activism (Dalgliesh, 2009) around ARA tools. A frontline practitioner (FP9) reflected, *"as professionals, they are not dealing with numbers; they are dealing with people who sat in front of them"*. It suggests that some practitioners have understood that ARA tools only give them a starting point for professional judgement. This is a momentous change for some practitioners: a transition from objectification to subjectification (Heller, 1996) and establishing a relatively new ethical discourse about ARA tools: The issue of labelling/categorising individuals by ARA apparatus. In that regard, a practitioner (FP4) mentioned this issue and explained, *"Staff have the opportunity to raise concerns in a staff survey. Whether they raise that in the staff survey, I don't know. But [Data team] do get information from staff in the Staff survey. But then the question is, what do they do with the information?* The findings suggest that the identification of this ethical challenge not only reveals that some frontline professionals have their own discourses of ARA ethics but also how their discourses have been pacified retrospectively by the ARA and actuarial data science.

5.4.4 Making ethical tools through ethical toolkits

The relevant literature on public administration, civil service, and criminal justice establishes a necessity that these organisations should be subject to scrutiny through academic research and parliamentary institutions (Cunneen, 2006). As described in the previous sections, the concerns around biased decision-making through algorithms exist. In order to respond to these challenges, the organisation has aimed to collaborate with external bodies (e.g., the Alan Turing institute: A research centre for AI and data science} to develop ethical frameworks/toolkits to ensure ethical ARA practices. In that regard, a senior manager (SM1) mentioned, "*I was one of a number of people across the department involved in different capacities with the Turing Institute because they were trying to come up with their kind of support. And to come up with an ethical framework that we could then apply so a set of values and a set of questions to ask."* A relevant strategy document (OD2) highlights, "*ethical values are designed to support, underwrite and motivate ethical conversations within and between teams (and organisations) by providing an accessible, common ground for thinking about moral scope of the societal and ethical impacts of data-driven technologies".* Indeed, by "data-driven", the organisation is referring to their algorithmic apparatuses for risk

assessment. The data shows that the organisation is keen to minimise ethical impacts of ARA tools, yet sees the solutions in relation to power/knowledge of expertise, including data experts. The external ethics bodies such as The Alan Turing institute are conventionally formed by university academics, industry experts, former professional and even political figures. These individuals possess the knowledge around data and computer science or may have worked around the development, adoption and impacts of AI technologies in various organisations. As a relevant document (OD1) indicates, their aim is to "*effectively and responsibly advance these [algorithmic] technologies in society"*. Thus, they conduct fieldwork and research to uncover and resolve ethical impacts from a technological perspective.

Some data scientists highlight that they have continuous collaborations with external bodies to uncover the ethical dimension of the ARA tool and to ensure that the tools are aligned with all legal requirements. A scientist (DS8) agreed, "*We've been doing some quite a lot of work with the Turing Institute and I guess our thinking about the ethical issues has expanded."* Despite the ongoing collaborations, the data team admits that there is still much to do. In that sense, a data scientist (DS3) said, "*Sometimes we've done this better. Sometimes we've done this less well."* One of the senior data scientists (DS5) reiterated the challenge of developing ethical guidelines for risk assessment. As they mentioned, *"Ethical AI is a very new term, especially in the past two to three years. So, it's a very new environment; a lot of people don't even know about it."* It highlights that integrating theoretical aspects of ethical AI with the existing practice is difficult since it is a new realm for many organisations, an issue highlighted by Ananny and Crawford (2018) as well.

As data indicates, senior managers and data scientists are keen for practitioners to be familiarised with the principles of transparency and explainability around algorithmic tools. In that regard, a senior manager (SM2) explained, "*We wanted staff to be able to understand [the algorithm's predictions] so that they were then making decisions.*" They have indicated that they want the frontline officer to understand how ARA prediction would affect an individual. The algorithmic outcomes should be explainable to the human user, a crucial issue of ethics which is argued by Adadi and Berrada (2018). Through collaboration with external institutes, the data team hopes to promote transparent and responsible use of algorithms and data in the criminal justice system. As it is enacted in a policy document (OD10), "*ethical principles provide actionable and operationalizable points of departure to help teams reflect upon and justify why the actions they have taken throughout their projects are, for example, bias-mitigating, non-discriminatory, and fair."* As it is shown, the data team necessitates rigorous understanding within the frontline environment in which the ARA predictions are understandable and justifiable by the end-users. This is an important and

positive measure that makes the ‘black box’ nature of algorithms more understandable and ethical (Geiger, 2017).

Such statements from the data scientists and senior management highlight an intensification of algorithmic power/knowledge (Hardy and Thomas, 2014). Science has already constituted its dominant discourse, which is the deployment of ARA technologies. The collaborations with external bodies to devise ethical framework only justifies the legitimacy of ARA practices. Furthermore, the existence of ethical frameworks makes any discourse from the frontline practitioners redundant. This is because the ethical frameworks are instruments developed by the scientists and experts for organisational stakeholders to raise awareness of any ethical challenge. Thus, it may create a blind trust amongst other actors that scientific discourse is the one and only ‘truth’ that helps overcoming potential ARA biases.

The people at the frontline, however, outlined that they have not been approached by data scientists or any external institutes for their input around the ethics of ARA tools. In that regard, a practitioner (FP8) said that “*at our level there was no involvement in that process yet.*” Although some frontline professionals have unearthed particular ethical dimensions within ARA tools, not many have been asked for their viewpoints. There was a unanimous response from practitioners (e.g., FPs: 5 & 8 14) that “*we’ve never been asked to”.* Furthermore, the frontline practitioners expressed their lack of awareness around the collaboration with external bodies and the development of ethical frameworks. It is interesting that although there are insightful views around ARA ethics from the frontline employees, the findings illustrate that those viewpoints are overshadowed or rather, marginalised by the dominance of expertise, i.e., that of the scientists and external bodies.

Through support from the senior management and the inputs from external bodies, scientific power/knowledge has been strengthened, which point to another important ethical implication. As my analysis explained, producing ethical frameworks means superiority of science power/knowledge over those of frontline staff around ethics. It suggests that

5.4.5 Overreliance on algorithmic predictions

One of the challenges in human-algorithm interactions is the extent to which humans perceive their intelligent counterparts as trustworthy or useful agents (Aoki, 2020). For instance, how much trust is a human end-user willing to put on algorithmic decisions and how the positive relationships between human and algorithm would impact organisational outcomes (De Cremer and McGuire, 2022). The literature has raised the issue of overreliance on algorithmic decision-making, highlighting it as an ethical concern and a

precursor to fundamental ethical debates in criminal justice organisations (Hartmann and Wenzelburger, 2021). As such, the organisational ethics literature signposts issues, such as algorithm aversion (Burton et al., 2020) and resistance (Pachidi et al., 2021) against algorithmic commodities that require further exploration. This section taps into these ethical debates. The findings suggest a polarity amongst the interviewed frontline employees. It shows a particular divide in viewpoints between the experienced employees, who have been in the job before the introduction of ARA, and relatively newer practitioners who joined the service after the ARA introduction.

According to the findings, adoption of ARA tools as a feature of risk assessment is embraced by many frontline professionals, especially by those who joined after the adoption of these tools. In that respect, a line manager (LM5) agreed, "*The IT system started getting better and better. But the actuarial data started getting more apparent and it [ARA tool] gives you a good indication, a good guide. It's a good guideline."* Many practitioners have expressed their enthusiasm and positive perceptions about the usage of ARA in the sense that the tools have become essential parts of their work practices. This group of frontline professionals believe that ARA tools are helping them to be more consistent and confident in their risk assessment decision-makings. A practitioner (FP8) reflected, "*I think generally the [ARA] tools are a really positive thing because we can't entirely rely on our own kind of dynamic professional judgment all of the time."* Another practitioner (FP6) highlighted that algorithmic predictions provide them with basic – but factual – information which is better than making "*guesses or gut feelings."* For some practitioners, algorithms have become pivotal instruments to exercise their power/knowledge. Yet, the relevant literature highlights how algorithms are able to control the working lives of subjects and make them believe that they are in control (Peeters, 2020; De Laat, 2019). A frontline practitioner (FP9) explained, *"I find predictive tools [risk of serious reoffence and sexual offence predictors] useful to stop my tracks and look at it again. They offer me another solution or offers me the opportunity to give that person [the offender] a chance to work on their problems"*. According to frontline members who advocate ARA, the tools offer them the chance to compare their judgements with the algorithm's output. Therefore, algorithmic tools offer some practitioners an ethical discourse that guarantees fairness for offenders but saves the practitioners from any upstream scrutiny.

Senior managers also reported that ARA technologies are introduced to ensure that officers' professional judgements are conducted fairly, based on the evidence and available information, and not driven by personal biases. A senior manager (SM2) explained, "*What we're trying to create here is the balance between the power of the tool to ensure consistency, efficiency, and fairness in the assessment."* Hence, the idea to augment

human-based risk assessment with algorithmic intelligence is to provide practitioners with a starting point (a baseline) towards assessing risks. The ARA implementation is initially designed to give practitioners directions and augment their risk assessments. However, other interviewees indicate that it is not always the case. According to a frontline professional (FP3), *"[ARA tools] can encourage ignorance and a lack of scrutiny as well, because you can become reliant and dependent on the tool*". Or another one (FP2) highlighted, *"I suppose we rely – really heavily – on those [ARA tools] rather than the professional judgment side of things.*" As such, relying mostly on algorithmic predictions is easier and safer for some frontline people to defend their judgement. Referring to this point, a line manager (SM6) mentioned, *"It's safer to agree with [the algorithm's] assessments rather than it is to disagree with them."* There have been circumstances where some frontline members have only taken into consideration the generated predictions of ARA. In contrast, another practitioner (FP2) suggested, *"You're still going to get some people [practitioners] who don't look at any of the dynamic factors and just agree with the tools, because that's the safe route to go down."* Hence, whilst ARA tools are utilised to give the practitioners a baseline or a starting point to risk assessment, some people solely use the algorithm-generated predictions. This point is argued in the AI/algorithm ethics literature as how algorithmic intelligence (agency) may threaten human agency, autonomy, and self-determination (Introna, 2016; Magalhães, 2018).

The above illustrates different discourses characterise ARA and raised ethical issues. Whilst some practitioners heavily rely on ARA, there are some able to exercise their agency and actively move away from algorithmic agential power (Peeters, 2020). Those have raised cautions around overreliance on algorithms and putting blind trust in generated outcomes and also illuminated their approaches to ensure algorithmic power is not replacing or circumventing their autonomy and/or agency. The way these practitioners ensure their autonomy and agency can be related to Foucauldian theories of *activism* and *subjectification* (McMurray et al., 2011; Foucault, 2019; Foucault, 1988). However, the practitioners' activism and resistance against ARA relate to the ethical issues in their profession. These subtle nuanced conducts are discussed in the following paragraphs.

According to the data, at the outset of the adoption and implementation of ARA tools, some frontline practitioners refused to remain reticent. Some frontline employees understood that algorithms might be a potential threat to their professionalism and job autonomy. In that regard, a senior manager (SM2) expressed, "*There has been quite deep-seated resistance in opposition amongst some of our staff group to these [ARA] tools."* In the early days of ARA utilisation, there were practitioners who were conscious of how algorithmic intelligence might result in voluntary or enforced redundancies. For instance, a practitioner (FP5)

suggested, "*these crazy scientific kinds of tools that appeared to be replacing professional judgment"*.

Findings indicate that the resistance amongst the opposing practitioners was more collective and was supported by the trade unions. A line manager (LM5) explained, *"We [trade union members] were very concerned about our professional judgement: We don't want it [ARA tools]. Stop it. Electronics. Electrics. No, we're not having them"*. However, the collective form of resistance has diminished over the years. As a senior manager (SM2) expressed, this is because *"People got used to it and accepted it [the ARA tools] as it sits alongside [their] professional judgment*". Or as reflected by the above line manager (LM5), "*We just tried to ignore it [ARA systems], and it didn't go away. Because progress is progress: IT, social media, etc. It's not going back"*. In other words, those practitioners have gradually come to understand the inevitability of algorithmic work practices and accepted it as a tool that helps them in their judgement.

Despite that many practitioners have adapted to ARA practice; a form of resistance is noticeable amongst some of the practitioners. For example, a frontline person (FP12) highlighted, "*Whilst we have to keep the [ARA generated] score into consideration when it comes to allocation, I can always use my professional judgment*." A line manager (LM1) reflected on how, "*we're able to professionally override the static [ARA] score".* These quotes suggest that some practitioners are able to exercise their agency and self-reflexivity (Raffnsøe et al., 2019) in their interactions with ARA tools. Rather than explicitly contesting the existence of ARA systems, these practitioners have implemented measures to ensure that their judgments are not influenced by algorithmic predictions. A practitioner (FP15) mentioned their awareness, *'obviously we're the people making the risk assessment at the end of the day, not the computer."* As such, these behaviours are in line with *subjectification* (Bergström and Knights, 2006). This process substantiates that these frontline professionals have an awareness around the issue of 'overreliance' and its ethical implications (Taddeo and Floridi, 2018). Hence, some frontline members have raised and lamented overreliance on ARA tools as an ethical concern, arguing that they are the final decision-makers, not the algorithms.

Findings also depict another form of awareness, which is spontaneous discussions or 'talks' within team members regarding ARA predictions. Some frontline practitioners are able to critically unearth, analyse and debate the factors that impact the ARA scores. For instance, a line manager (LM4) explained, *"Some [frontline] people were really analysing that [ARA prediction/score] in their mind and able to think it better through."* The line managers have found these ad hoc conversations are most valuable since practitioners can

reflect on the performance of algorithms and are able to argue against it if necessary. A practitioner (FP12) highlighted this: "*We have to complain to our line manager, who will then convey the messages later on to other people."* It indicates that some practitioners can voice their concerns around algorithms. Although the previous format of collective resistance – via trade unions – is less likely nowadays, novel forms of awareness and 'resistance' have emerged that entail professional oversight as well as internal discussions amongst line managers and team members.

Findings show a number of practitioners prefer not to override/change the algorithm's predictions since it is safer to agree with it, particularly when a prediction has tagged an offender as medium to high risk. An experienced practitioner (FP2) illuminated this issue clearly: "*If someone got a medium ARA score, very, very few colleagues would ever put a low risk of serious offence. **I do** …But most of them would be **too afraid** to have a low risk of serious offence score if the ARA score was medium*". Such arguments indicate that while practitioners feel safer relying on algorithms [despite the potential flaws/biases], there are some who oppose algorithms. These particular groups of practitioners are aware of the risks of bias/prejudice in algorithmic predictions and refuse their conduct to be affected by an algorithm (De Laat, 2019). Such statements reflect the arguments by Weiskopf and Hansen (2023) that there is space for ethical practice, as some practitioners have become aware of and acted to minimise the impact of algorithmic dominance. Although it seems there is an intensification of scientific-algorithmic power/knowledge in this organisation (Hardy and Thomas, 2014), my findings unwrap a discourse of ethics formed through organisational employees' awareness, activism, and nuanced resistances. The practitioners' actions including trade union collective bargaining, talks/debates around ARA predictions and professionally overriding are examples of this novel form ethical discourse within algorithmic work environment.

5.5 Conclusion: Human professional judgement or the use of ARA

As it was shown previously, the possibility of *inconsistent risk assessments by practitioners* was invoked as a pretext to render ARA tools and better control [and regulate] human professional judgments (Bucher et al., 2021). A senior manager (SM2) explained, *"the sweet spot for me is that balance between using systematized [ARA] tools and data and that predictive capabilities alongside human professional judgment".* The same senior manager (SM2) highlighted, *"[The ARA tools] are impactful, and they are only scratching the surface since the service has so much data. And there is still lots of potential*." This approach is, of course, directed towards ethical service delivery. As another senior manager (SM1)

emphasised, "*We want to make sure that [the intervention] is delivered to the people they're right for."* The senior leadership emphasises that the service should continue to grow the use of algorithmic predictors not at the expense of replacing human cognition, but to supplement or augment it. However, an ethical takeaway discovered through data is the marginalisation and abandonment of practitioner end-users. Initially, practitioners were excluded from the decision-making processes to utilise ARA. The organisation has predominantly favoured data science power/knowledge over practitioners' experiences.

Many frontline practitioners and line managers [senior staff], however, offered a different perspective around their work experiences, specifically relating to the importance of ethics in the decision-making for their clients. This notion has indeed benefited this research with a novel insight on algorithm ethics. As such, some practitioners agreed that there are ethical issues such as data limitations and lack of accurate variables in ARA tools. A practitioner (FP6) agreed, "*Algorithms can never be a panacea to improve decision-making*". Another practitioner (FP3) reiterated, "*The general creation and adoption of an algorithm is not the answer to the problem. But rather*, *it is something to have in the toolbox.*" Such statements indicate that for some practitioners, an algorithm is solely an assistive tool, rather than a conclusive decision-maker. A line manager (LM7) explained, "*The only thing that worries me is that they [ARA tools] take away the way a human thinks; when you're talking about crime and victims, do you think a computer can really empathise with that and sympathise with that?"* These actors have expressed their awareness and knowledge, arguing that the dominance of ARA power/knowledge might result in the deterioration of human agency, control, and professionalism, as argued by Magalhães (2018). According to these practitioners, there is a risk that their subjectivity and agency may be lost within the power/knowledge of algorithmic science.

Many frontline employees have understood and praised the benefits of ARA commodities. Many practitioners listed these benefits as "*systematic information on offenders"* (FP17), "*statistical/evidence-based risk assessment"* (FP8) and *"standardisation framework for decision-making"* (FP1). Yet, some still believe in the supremacy of human intelligence over algorithm technologies. In that regard, a practitioner (FP40 mentioned, *"Algorithm simply lack emotions, feelings, grief*." A line manager (LM40 highlights this deficiency, arguing, *'This comes into the point about ethics. I would like people [frontline staff] to be scrutinizing these information [ARA outputs]".* This again highlights the extent of practitioners' activism and awareness around the flaws of algorithms and the discursive shift towards 'ethical conduct'.

A number of practitioners have rallied against ARA tools by arguing that human professional judgments must come first and cannot [should never] be removed from the virtue of the criminal justice system. These practitioners understand the potential ethical flaws of ARA yet have gone above and beyond such technological discourses. Some even argue that the criminal justice service should be restructured radically and move away from the bureaucratic work. Reflecting on this issue, a practitioner (FP5) highlighted, *"I think we're a service without its soul that has totally lost its way. And there is a complete lack of collaboration between practitioners and those in charge of the service"*. Remarks as such suggest that although there are signs of bureaucratic control over the working lives of practitioners (Hodgson, 2004) via scientific-algorithmic power/knowledge. Some are aware of this disciplinary subjection and are able to refuse/resist it. The findings indicate that not only are some practitioners aware of the ethical ramifications in algorithms (e.g., bias), but they also understand how scientific power/knowledge relations are systematised to administrate and control their working practices (Hardy and Thomas, 2014). As such, activism through subjectification (Heller, 1996) and subtle workplace resistance (Alakavuklar and Alamgir, 2018) have mainly been focussed on ethical practices and their concern to do justice in their decision-makings for their clients. Many practitioners are able to exercise agency, subjectivity, and activism to avoid or minimise the impact of algorithmic work practice, illuminating instances for ethical conduct.

In this chapter, I highlighted the divisive discourses that reflect the ethical dimensions of the adoption and implementation of algorithms for risk assessment practices in a criminal justice organisation. It has explored the intensification of a particular power/knowledge discourse that stemmed from data science expertise, which led to the introduction of ARA tools. Initially, the processes through which scientific power/knowledge relations marginalised frontline practitioners' knowledge were justified through the notion that ARA had less bias, and practitioners might be biased. ARA tools were held up as directing work towards more 'consistency'. Promises such as bias-free human judgement, enhanced rehabilitation and protection of community, data science, and algorithms were seen as fundamentally superior as they strengthened the working lives of organisational employees. Despite the intensified dynamics of ARA power/knowledge, the frontline practitioners have been able to critically scrutinise the algorithmic tools and exercise their agency and subjectification in these power relations. The findings shed light on the nuanced 'ethical conduct' of some practitioners, including talks/debates on ARA performance, danger of overreliance, discard, change, or professional override of algorithmic predictions. I explained that such conduct can be considered as instance for space for ethics (Weiskopf and Hansen, 2023) and a different discourse that was developing within the dominance of the scientific

power/knowledge regime. These instances of ethical activism are mechanisms used by some employees to offer a more ethical service and minimise the impacts of the ARA scientific power/knowledge. In the next chapter, I discuss the nuances of competing ethical discourses through a Foucauldian lens and use them to answer the research questions.

CHAPTER 6: Discussion

6.1 Introduction

In this chapter, I discuss the findings highlighted in the previous chapter in relation to the existing literature. The theoretical and managerial contributions, limitations, and recommendations for future research are also presented. In chapter 1, I proposed and asked these research questions: Firstly, I asked, *what are the dominant ethical discourses around the deployment of algorithms from the perspective of key organisational actors?* And secondly, I proposed, *how do organisational actors influence the ethical discourses of algorithms, and to what extent do they change or challenge their work experience with these tools?* In order to address these research questions, I have adopted a Foucauldian perspective, incorporating his critical look on discourse, governmentality and resistance/ethics. In this chapter, I use this Foucauldian lens as an analytical lens to make interpretations from findings, answering the research questions.

As suggested in chapter 5, the targeted criminal justice organisation is structured on the principles of enhanced rehabilitation and protection of the community. These two values have been the pretexts for the organisation to introduce new practices, methods and technologies. Via utilising algorithms, the organisation aims to ensure that the offenders will receive the most suitable rehabilitative interventions as well as guaranteeing the safety of the community (Hartmann and Wenzelburger, 2021). Due to the nature of the work practices, many criminal justice settings have been criticised as institutions with potential for human bias, and palpable instances of racism or discrimination (Davis, 1996; Cunneen, 2006). Thus, this service has been keen to adopt work commodities which will contribute towards an unbiased impartial risk assessment. ARA technologies have been chosen as a superior tool to ensure the frontline practitioners' unconscious bias is minimised. However, the adoption of algorithms to fulfil the ARA's promises highlights some ethical nuances that can affect not only organisational employees but also their clients. In the following section, I will highlight the novel findings that emerged through data analysis in relation to the research questions. Subsequently, I illustrate their implications for the existing literature, explaining how these findings extend the knowledge on AI/algorithm and the ethical aspects.

6.2 A summary of key findings

In chapter 5, I demonstrated that there are several ethical ramifications regarding the adoption and implementation of ARA predictive tools. These ethical ramifications are not *per se* due to the nature of algorithms and their internal opaque processing (Geiger, 2017);

rather, these ethical issues have emerged as people's (i.e., frontline practitioners) discourses were excluded from the key decision-making processes at early stages of introducing the ARA tools. As findings indicated, the frontline practitioners were keen to be discussed and share their viewpoints on ARA practice; however, according to the interviewees, no significant discussion has taken place. The lack of inclusion and marginalisation of frontline practitioners' voices has resulted in issues such as mistrust towards algorithmic predictions (Russo et al., 2024) as the practitioner end-users often criticise the ARA predictions. The findings highlight that whilst ethical practice, for data scientists and senior leadership, is envisaged via enhancing human cognition through utilisation of ARA, the employees at the frontline illustrate ethical practice in a different manner. Frontline practitioners were keen to be involved in the processes of design, development, and implementation of ARA; they wanted to have a 'say' in the processes/decisions that led to the adoption of ARA tools. The exclusion of frontline employees' voices is considered an ethical question, whilst it was not raised as an ethical matter in other stakeholders' eyes, including data scientists and senior leaders.

The research findings, however, illustrate a shift in the discourses of many organisational actors regarding their concerns on the efficiency and practicality of ARA technologies. As such, a number of practitioners have raised questions on how accurately ARA tools can predict risk and whether the tools are taking into consideration all factors, variables, and other relevant information prior to predicting a risk about an individual. Additionally, some frontline practitioners have criticised the lack of effective collaboration between teams/departments within the criminal justice organisation in relation to topic ARA tools and their flaws. These issues were identified and raised as ethical matters of algorithmic tools, stemming from, particularly, the frontline practitioners' work experiences and awareness. Their point is that although the adoption of ARA tools is a significant step in tackling human biases and making more ethical decisions, there are still flaws associated with algorithmic predictions. However, due to ineffective collaboration between data scientists and frontline teams, the identified ethical issues are not conveyed and have remained unresolved.

The last key finding highlights particular ethical nuances that emerged through discourses of frontline practitioners. As ARA tools have been a well-established commodity with the criminal justice service, many frontline employees have been able to identify nuances of ethics related to algorithms. These ethical nuances explain how ARA outputs may lead to categorising or labelling individuals. Furthermore, these frontline employees have taken measures to minimise the impacts of ARA tools, incorporating more human input in the assessment of risk. So, by adding more human elements, these practitioners have

engaged in ethical conducts, ensuring that they are offering an ethical transparent risk assessment. Although the existence of ARA tools was fundamentally designed to ethically regulate and normalise the work practices of frontline practitioners, findings indicate that many practitioners do not let ARA tools steer their cognition and/or direct their decisions; rather, they are the ones who ensure that algorithm predictions are free from bias or prejudices by professionally overriding or disregarding the predictions. These practitioners deem such conducts necessary because they have understood that the algorithm's generated output is not in line with ethical values or is not fair. The highlighted discourses and actions of this particular group of practitioners are unique and unprecedented as they illustrate the extent to which human end-users can/will exercise their own power and agency (De Laat et al., 2020) in human-algorithm work interaction. In the following sections, I will look through these key findings via using a Foucauldian lens, arguing how my findings can shed further light on the blind spots of literature on organisational ethics as well as AI/algorithm ethics.

6.3 Theoretical contributions to Organisational Ethics literature

6.3.1 Governmentality and ethical management

Foucauldian theory of power/knowledge problematises 'power' as discursive systems that are not possessed, granted, or taken back, but rather can be 'exercised' in a productive way, informing new ways of behaviour (Foucault, 1977). In other words, power can be exercised through means and actions. His understanding of power stands against concepts of sovereign power that can be inherited or possessed by individuals (Knights, 2002). It also moves away from structuralism/Marxism power that depicts power as an economical force of production (Nigam, 1996). In line with the concern around the notion of power, Foucault has focused on how neoliberal governments use technologies, practices, and rationalities to exercise self-regulation of individuals (Raffnsøe et al., 2019). He uses the term 'governmentality' to refer to those technologies, including discourses and rationalities as strategies used by the governments to influence individuals to self-govern themselves (Weiskopf and Hansen, 2023). Foucauldian governmentality can be informative to understand how neoliberal governance can impact a government-led criminal justice organisation by introducing novel algorithmic tools, justified through a discourse that prioritises the knowledge of data experts over the employees' knowledge. Foucault's lens of discourse enables us to understand why the employees either embrace or are hesitant towards the new algorithms, and how this nuanced resistance slowly manifests through a

discourse cornering ethics and the impacts of the expertise power/knowledge and technologies on their clients (Shaw and Scully, 2023).

As it was shown in the previous chapter, the criminal justice service in this study has adopted algorithms to ensure risk assessment practices are consistent, transparent, and ethical. To do so, the organisation has invoked the expertise power/knowledge of data scientists to design and deploy ARA tools to augment the working lives of frontline practitioners. However, as data illustrated, algorithmic transformations took place in the organisation with only limited discussions and negotiations with the practitioners. Furthermore, in terms of decision-making for algorithm implementations, findings showed the marginalisation of the frontline practitioners in relation to many decision-making processes. Scholars have discussed the ethical issues surrounding the practice of governmentality particularly in organisation and management studies. For instance, Cludts (1999) advocates participation/inclusion of employees in defining shared values (such as ethics) of the organisation. In his theoretical work, ethical participation is seen as stakeholders engage in dialogues to create a consensus based on discourses.

The art of governmentality being exercised via algorithm technologies has a number of ethical ramifications. Such ethical ramifications *per se* are not linked to the nature of algorithms or their internal processing, but to the rationalities and strategies used by the government to steer such technologies into the work practices. Firstly, the adoption of algorithms shows the dominance of data science power/knowledge as an accepted 'truth' towards ethical practice. Hence, the utilisation of the scientific method to assess individuals and predict the likelihood of an offence occurrence has become a pervasive power/knowledge discourse (Townley, 1993). Foucauldian power/knowledge explains the extent to which power relations derive from subjects' conformity to discursive practices, which subsequently makes controlling the population easier. Moreover, it explains how people become subjects of knowledge that further tightens the exercise of power (Clegg et al., 2006).

In the previous chapter, I highlighted how the organisation deemed algorithmic technologies to offer more consistent predictions compared to frontline practitioners' judgements. Thus, the practitioners became the end-users of algorithms and objects for data science to evaluate the performance [knowledge] of algorithmic intelligence. The scientific power/knowledge, thereby, has gained dominance over human-based risk assessment to enact that ethical [consistent] risk assessment through human-algorithm collaborations. Huda's (2019) study suggests that ethical/moral deployment of any technological tool is associated with whether and how that technology is used to foster the societal and

organisational benefits as well as the professional development of humans. The research shows that the utilisation of ARA tools has contributed towards achieving the organisational goals with partial societal benefits. However, the issue of ethics in relation to the client's sentencing has become the thorn in the criminal justice behaviour. Similarly, Chatterjee et al. (2023) research argues that ethicality and morality in the adoption of technology are very much related to the extent to which the organisation endeavours in the governance of that technology. In that regard, although findings of this study suggest that data scientists are working towards better governance and ethicality of algorithms, the initial exercise of governmentality relatively overlooked frontline professionals' discourses. This is in line with previous research that highlights the risk of loss of human agency, self-reflexivity and autonomy due to the exercise of governmentality via algorithms (Introna, 2016; De Laat, 2019).

There is research that has focused on the inclusivity of human end-users in the design and delivery stages of introducing a technology. Robillard et al. (2018), for example, identify the engagement and involvement of users [beneficiaries] as ethical tenets of adopting any technology. They argue that the design of a technology should be user-led, participatory, and based on their needs so that it can be deemed ethical and just. But it seems the existing literature has not considered the impact and dominance of scientific discourses that drive decisions and enact ethical guidelines. The research findings illustrate that the decisions and processes to implement ARA tools were predominantly steered by data scientists who were also the sole designers of the tools. This is in line with the Foucauldian theory of governmentality (Moisander et al., 2018), but the ethical side of governmentality and scientific power/knowledge techniques (Raffnsøe et al., 2019) are yet to be challenged.

Previous research has investigated the other form of governmentality via algorithms, which is 'surveillance' (e.g., Zuboff, 2019; Roberts, 2019; Newlands, 2021). This strand of research sheds light on neoliberal governance seeking to better control and administrate the lives of people via constant algorithmic surveillance and highlights the violation of ethical values such as privacy of citizens (Murphy, 2017). However, there is limited insight on other types of algorithmic governmentality and relevant ethical discourses. My findings indicate an exercise of disciplinary tactics (i.e., via ARA practice) to create regulated individuals without imposing any repression, prohibition, or coercion (Bergström and Knights, 2006). Algorithms have become a *power exercise* instrument to better regulate and normalise practitioners' working lives at the criminal justice service. These tools subtly control and regulate the working lives of employees through dominant discourse, which can damage 'ethical practice' and transform human-based practices (Flyvbjerg, 1998). The algorithm tools' predictions are

based on factual evidence and computer intelligence. Hence, the data science as well as organisation have been able to constitute the tools within frontline practitioners' professional judgement. In that regard, I argue that ethical risk assessment, from data science and senior management perspectives, is envisaged through utilising algorithms, augmenting practitioners' cognitions and decision-making capabilities. As such, this study joins the discussions raised by Hartmann and Wenzelburger (2021) in the sense that algorithms are able to provide more impartial, non-discriminatory predictions for criminal justice practices.

However, the Foucauldian lens of governmentality helps to explain the ethical dimension of algorithm adoption in relation to the ethical conduct in the organisation (Chye Koh and Boo, 2004). It draws attention to techniques, strategies, and political agendas that have constituted algorithms as a discourse to ensure practitioners' compliance with the ARA tools. This compliance challenges organisational justice and employee perception of fairness (Crawshaw, 2006; Törnroos et al., 2019; Laundon et al., 2019). Organisation ethics literature is predominantly associated with the workforces' perception of fairness as well as employee-employer relationships. Moreover, previous research considers *procedural justice* and discusses the extent to which organisational actors perceive the workplace procedures as fair and free from bias (Roberts and Herrington, 2013).

The impact of imposed unfair policies and practices on job satisfaction (George and Wallio, 2017) and workforce turnover rate (Pieters, 2018) has been studied previously. However, there is only limited discussion on the impact of governmentality within organisational ethics literature (Raffnsøe et al., 2019). The research findings emphasise governmentality as an ethical issue, as practitioner end-users were not appropriately involved in the decision-making, design and development processes of ARA tools. It was raised by many experienced practitioners who were concerned about the algorithmic outcomes affecting their clients and a different discourse emerged. Foucault explains discourse as a system in which different knowledges and power relations compete (Smart, 1992). The discourse of the ARA systems embedded in a criminal justice organisation was challenged by some experienced practitioners on the basis that ARA is not nuanced for ethical decision-making. Hence dominant discourses are contested and aimed to inform some change through power, knowledge and action (Smart,1992). My findings shed light on the influence of governmentality via algorithmic discourse that marginalises other discourses, including practitioner end-users from the adoption processes.

Foucault notes the tensions between the negative impacts of technology, seen as a form of social and political control that should be subject to critique, and positive impacts that can offer solutions to previously unacknowledged limitations (Behrent, 2013). He also uses

the term ‘technology’ to denote procedures that manifest as political technologies that foster discipline and regulation. Discipline provides means for control to be exerted over an individual’s conduct, aptitudes, performance, and capacities (Foucault, 1972). In that sense, ARA disciplined practitioners and presented a discourse that challenged their capabilities as inferior and suggested that they were biased in their decision-making. Foucault explained the concepts of objectification and subjectification as a process of resistance. Objectification describes the process by which people become the objects of discourse as a result of pressure to obey and submit to expert power/knowledge (Heller, 1996). According to Foucault (2000), the discursive process of objectification, the prevailing dominant discourse, is internalised, and people confront their own thoughts and ways of being. This can manifest in the objectification of the self by oneself, others, or technology (Bergen and Verbeek, 2021). Subjectification occurs when people become resistant to practices deemed desirable by the dominant discourse (Foucault, 1983). Thus, people are simultaneously objectified and subjectified by discourses. Objectification is based on processes of power relations (power/knowledge) that can impact individuals or society as a whole (Khan and MacEachen, 2021). My findings illuminate how some practitioners confronted this process and resisted the dominant discourse through their knowledge of ethics in their professional role and their concern for the decision making that impacted their clients. Other practitioners, particularly the young (less experienced) practitioners, embraced ARA and saw it as tool that took the responsibility for the decision-making in sentencing. This is because the dominant discourse was advocating that ARA had no bias in the decision-making, whereas the practitioners were biased in their decision-making.

Foucault tells us that people are products of discourses (Alvesson and Karreman, 2000). For Foucault, objectification via technologies and state apparatus gives the individual (subjects) the illusion of choice in how they construct themselves (Bevir, 1999) People are, therefore, objectively managed by the mastery of the discourse (Townley, 1998) which, in the case of ARA implementation, is strengthened by trust in the power/knowledge of both the government and technology. As products of discourses, subjects concurrently create, and are created by, discourses (Clegg and Dunkerley, 1980; Foucault, 1982). Subjectification via technological power becomes a form of power that makes individuals subjects and submits them to others (Foucault, 1983) whilst recording and reproducing knowledge and apparatus to formulate social hegemonies (Nola, 1998; Foucault, 1972). This process is the result of the interactions between a subject’s agency, their political and social affinities, and the organisational discourses used to construct realities and meanings (Foucault, 1982; Hildebrand-Nilshon et al., 2001).

Through the process of subjectification, people can transform discourses as they become more self-aware of their power/knowledge and agency and resist their objectification and subjectification (Heller, 1996). This requires using knowledge to develop an alternative discourse (Clegg, 1998), and some form of challenge towards this oppressive discipline (Goodwin, 2019). The knowledge and capacity to challenge a dominant discourse may not be readily available in contexts where expert knowledge is justified by the state and the organisation. However, resistance gathers momentum once it is being discussed. This research can inform and extends organisational ethics literature in a sense that the exercise of governmentality aims to make *self-disciplined* employee practitioners. It is carried out as the expert's knowledge highlights the benefits of algorithmic work, ensuring frontline practitioners' complacency in their interactions with algorithmic agents. The ethical issue, however, is the extent to which novel algorithmic governmentality marginalises and excludes other discourses, such as practitioners in ARA practices, which are important and informative.

6.3.2 Discursive Power/Knowledge

Previous research on algorithms and governmentality discusses the extent to which individuals become the sources of data, which, subsequently empowers the algorithmic apparatus of governmentality (De Vaujany et al., 2021; Leonardi and Treem, 2020; Walker et al., 2021; Zuboff, 2019). This strand of research sheds light on growing concerns on how people's behaviours may be influenced and/or manipulated by AI and other intelligent agents. Moreover, this research highlights organisational risks such as ambiguous transparency (Leonardi and Treem, 2020) or disconnection of employees from the organisation due to extensive insistence on data-driven technologies (Hafermalz, 2021). There are also studies providing evidence on how socio-political power is being exercised by algorithmic technologies in different contexts from neoliberal governmentality doctrines (Königs, 2020). However, there is a need to better problematise and theorise human behaviours who interact with data-driven tools and to better understand the conditions and consequences of such technologies (Leonardi, 2021).

As findings demonstrate, ARA tools use variables in data such as age, gender, and ethnicity to process offence cases. These variables are collected from different sources, which depicts only a limited image of individuals' characteristics and may lead to biased risk assessment according to Tsamados et al. (2022). On the one hand, senior leaders and data experts do not raise this issue as an ethical matter, which is explainable as the aim of exercising power/knowledge is to constitute disciplined individuals through creative

strategies, tactics and dynamics (Foucault, 2020a). As such, data-driven technologies, together with the organisation's inscribed ethos of *enhanced rehabilitation* and *community,* are seen as the solution to fulfil the organisation's promises. On the other hand, a group of practitioners had the knowledge to challenge this discourse through their lived experiences. As shown in the findings, the art of governmentality has been questioned/challenged by some practitioners as they have understood and criticised the mentioned ethical flaws in the ARA practices.

Foucault (1982) explains this as the power/knowledge mechanisms that work in creative ways, subtly constituting individuals as self-disciplined who conform to established practices. Although some of the practitioners have raised their concerns, others show their compliance with predictions of the tools. Hardy and Thomas (2014) indicate that there is potential for power/knowledge strategies to intensify and become perpetuated as a dominant 'truth' in organisational processes. This argument is evident as many practitioners feel safer using ARA tools in their judgements.

The criminal justice organisation has aimed to minimise the risk of human biases in their risk assessment. Algorithmic decision-making is generally known to offer impartial outputs and manifest lower degrees of biases/prejudices compared to human decision-making (Charlwood and Guenole, 2022). The flaws surrounding human judgements are explored and have become evident in recent research (Howard et al., 2020). However, algorithms are also subject to the risk of bias (Diakopoulos and Koliska, 2017). The impact of human biases, stereotypical thoughts, and prejudices is evident in organisational decision-making (Jones and Roelofsma, 2000). Subsequently, the introduction of ARA tools was aimed at overcoming potential human biases in the risk assessment and providing factual evidence for any scrutiny from public or political institutions.

The findings raised an ethical question in the sense that the utilisation of ARA technologies might undermine the value of human-based professional judgement. Due to statistical reasoning and computational processing, ARA tools are deemed to perform better, offering more impartial risk assessment, free of human biases and prejudices. Therefore, it seems that these tools have circumvented human professionalism. Many practitioners seemed to have accepted that their judgments may be biased; hence, they ought to use ARA tools. As Foucault argues, governmentality calls upon people to act and gives them 'some' authority or freedom (Raffnsøe et al., 2019), whilst, actually, the people are governing themselves to foster neoliberalism (McKinlay and Starkey, 1998). It is in line with how the dynamics of power/knowledge and governmentality work in organisations according to Foucault (2020e): Neoliberal governance implements technologies in order to show that they

have concerns around societal issues and intend to rectify them (Weiskopf and Hansen, 2023).

Foucault does not limit the forms of governmentality to only political state governments. Rather, he suggests that this regime of knowledge exists within other forms of organisation such as the administration of families, schools, and medical clinics (Raffnsøe et al., 2019). The criminal justice service, in this research, has introduced ARA tools to respond to the ethical questions surrounding the human-based risk assessment. However, this strategy has established another challenge of ethics related to the practitioners' professionalism. The criminal justice organisation exercises an ideology that assumes the algorithmic technologies can offer fewer biases and foster consistency in decision-making. This ideology is based on reciprocal ethical norms and values of the organisation and the society. However, their ideology is – as conceptualised by Foucault (2020b) – a 'discursive governmentality' that not only might make frontline professionals marginalised, but also lead to further intensification of scientific data-driven rationalities. These ethical dilemmas predominantly reside in modern state neoliberalism ideology and constitute a power/knowledge sovereignty for the use of ARA technologies. As such, the research findings further expand organisational ethics literature, specifically around employees' dignity and self-determination (Gibson et al., 2022; Lucas, 2015; Sayer, 2007), by unearthing how 'discursive governmentality' via algorithms is capable of dehumanising work practices. This research establishes a single truth that other organisational actors would consciously follow.

6.4 Theoretical contributions to AI/algorithms ethics literature

6.4.1 Self-reflexivity and concern on the use of ARA technology

More recently, it has been argued that the trustworthiness of algorithms is a key factor in human-algorithm interactions (Aoki, 2020; Dwivedi et al., 2021; Durán and Jongsma, 2021; Floridi, 2019; Roßmann et al., 2018). This strand of literature outlines that an algorithm's trustworthiness is directly linked to important concepts such as accountability and transparency of algorithms used in work practices (Okamura and Yamada, 2020). Foucault has conceptualised ethics as an individual's self-awareness on the issue of disciplinary power/knowledge (Foucault, 1988). As findings indicated, the introduction of ARA tools was received with mixed views, including scepticism and doubts amongst some of the experienced practitioners because they saw algorithms as a replacement for their professionalism. When ARA was established, the experienced practitioners were aware of the potential disruptions that ARA could cause in their professional judgment. Whilst the data

scientists and senior leadership initially aim to portray ARA tools as advantageous tools, those opposing practitioners raised doubts on the tools' effectiveness. They raised their concerns on how tools might overlook many other crucial factors – particularly the psychological issues – associated with the prediction of risk. Hence, some practitioners conveyed their concerns through existing channels, such as workforce unions.

As discussed above, the marginalisation of practitioners' voices has not gained total dominance over practitioners' power/knowledge. This research is conducted years after the deployment of algorithms in the organisation. But there is still palpable criticism that highlights degrees of mistrust amongst some practitioners. These concerns can be explained through Foucault's theories of subjectification (Heller, 1996), self-reflexivity and awareness (Crane et al., 2008). The self-reflexivity, or self-awareness is the journey that practitioners may take to become aware of being objectified to a disciplinary mechanism of power (Crane et al., 2008). These concerns are raised by some practitioners in relation to ARA commodities and awareness and subjectivity in the algorithmic power/knowledge regimes. There is considerable mistrust amongst some practitioners in terms of potential flaws of algorithms. Thus, frontline professionals have been able to reveal the ethical downsides in the algorithmic tools and have constituted themselves as 'ethical subjects' according to Foucault (2020c).

There are many studies that focus on the issue of (mis)trust and algorithm aversion (e.g., Dietvorst et al., 2015; Burton et al., 2020; Heßler et al., 2022). These studies predominantly depict underlying reasons within the technology that result in human end-users' mistrust of algorithms. For instance, some research sheds light on the perception of an algorithm's cognitive inferiority compared to algorithms' that lead to the exhibition of aversion or mistrust towards their intelligent counterparts (Burton et al., 2020). Although such studies have considered factors such as algorithm bias or the dehumanising nature of algorithms as precursors for mistrust of algorithms, their arguments remain fairly hypothetical in the need of more evidence from organisational settings. Furthermore, it seems that ethical aspects are not major points of debates in such studies, and there is only limited attention to this pivotal element (Jauernig et al., 2022). Which raises the question: to what extent do ethical dilemmas of algorithms trigger concerns amongst human end-users and cause further aversion against algorithms? By building on Foucault's work on self-reflexivity, self-awareness and subjectification (Heller, 1996), the possibility of how some frontline professionals – as the end-users of ARA tools – identify the potential flaws of such tools can be explained. Any existing mistrust amongst these practitioners' stems from their awareness and concerns around the extensive utilisation ARA practice.

This study argues that ethical practice is not solely about augmenting human decision-making by algorithms. It also involves mistrust, expressions of concerns and communication with other organisational actors to challenge the existence of ethical algorithmic practice (Weiskopf and Hansen, 2023). As such, many practitioners have constituted themselves as ethical subjects. Foucault uses the term 'ethical subjects' in his work about individuals who are aware of power/knowledge orthodoxies exercised over them (Crane et al., 2008; Gardiner, 1996). Therefore, this study adds to the existing literature on human-algorithm interactions by pointing out the nuances of mistrust, concerns, and debates around the utilisation of algorithms. By using a Foucauldian lens, not only this research provides empirical evidence on the issue of mistrust in human-algorithm interactions literature, but also better theorise ethics in relation to actor's mistrust, concerns, and doubts in the algorithmic work practices.

6.4.2 Concerns on lack of collaboration and disconnect

Public administration organisations are known for their top-down hierarchical structures (Hill and Lynn, 2004). One of the issues associated with these top-down bureaucratic organisations is the possibility for ineffective communications between different levels of organisation (Meijer, 2008). Several challenges have emerged due to ineffective communications in this criminal justice service. Firstly, it seems that the deployment of ARA tools, initially, was not properly explained, communicated, and promoted by the lead decision-makers. Secondly, interactions between the frontline practitioners and data experts regarding ARA tools have remained minimal and limited. Also, due to a lack of proper communication and interaction, some tensions and challenges have emerged.

Previous research has tended to understand organisational tensions through theories such as 'the theory of paradox' (Lewis, 2000). The research surrounding the theory of paradox offers an alternative approach to 'contingency theory' (Schoonhoven, 1981) for better understanding and managing the organisational tensions and issues (Smith and Lewis, 2011). It suggests that tensions and challenges can, in fact, be beneficial as organisational actors with different perspectives coexist. Put differently, a paradox lens envisages the solution to tensions in simultaneous engagements between the two poles of the tension. Studies that have subscribed to the theory of paradox provide insights to better understand and solve organisational phenomena that stem from tensions (e.g., Carmine and De Marchi, 2023; Miron-Spektor et al., 2018; Ozanne et al., 2016). Yet, the theory of paradox offers limited avenues to understand organisational actors' discourses, particularly, actors' agency and subjectivity. A Foucauldian perspective, however, includes people's

subjectivity, agency, discourse, and activism as integral components (Raffnsøe et al., 2019). Aligning the theory of paradox with Foucault's work of human subjectification (Heller, 1996) offers an opportunity to understand the tensions and challenges amongst actors in light of algorithmic work practices. Foucault shows how discourse regulates individuals in a way to constitute them as instruments of power/knowledge (McKinlay and Starkey, 1998), whilst the theory of paradox incorporates subjects' agency and self-reflexivity to offer solutions to the tensions of algorithmic work.

The discussion, so far, has explained that the data scientists' insistence on ARA tools acts as a discursive instrument for excising scientific power/knowledge orthodoxy. Continuous support and endorsement from senior management further perpetuates the algorithmic apparatus in the criminal justice setting (Pachidi et al., 2021). However, this study highlights a 'disconnect paradox' (Leonardi and Treem, 2020) between different organisational employees. This paradox has emerged as some practitioners felt left out from ARA work practices, particularly as the technology seems not to resonate efficiently within their work. The lack of a clear explanation about the ARA tools have undermined practitioners' autonomy and have disconnected them from their organisation's environments. This is in line with Leonardi and Treem's (2020) discussion about how a technology that is intended to create consistency and harmony amongst organisational actors may lead to their disconnection, causing tensions/paradoxes as they seek effective connectivity with their organisation. Foucault's work highlights the subject's journey on becoming self-aware around the disciplinary mechanism of power (Knights, 2002). However, there is critique on his work around lack of dialogue (Gardiner, 1996) and human agency as discourse (Caldwell, 2007). Although his later work considers the modalities through which subjects can become the ethical, there is a tendency to focus on individualism or the self (Foucault, 1980) and refusal of elements such as dialogue between all reflexive agents (Flyvbjerg, 1998).

Foucault's conceptualisation of the ethical subject makes theoretical contributions as it is juxtaposed with the theory of paradox (Lewis, 2000). Foucauldian revelations on self-reflexivity and the ethical subjects illuminate how some practitioners become aware of the disciplinary power of algorithmic work. These practitioners expressed their concerns on the facilitation of ARA and have sought to reclaim and retain control over their work practices to become 'known' and constitute themselves as 'ethical subjects.' Simultaneously, the theory of paradox helps to better understand how organisational actors' endeavours to regain control can be used to find appropriate means of communication and connectivity between, for instance, data science and frontline members. This study, thereby, argues that the combination of Foucauldian lens with the theory of paradox would benefit the literature of

AI/algorithm ethics in two ways: First, applying Foucauldian lens would help to profoundly identify the tensions, challenges, and concerns of organisational actors to take back control of their work in algorithmic work practices. Second, the paradox theory dimension can help to enhance organisational strategies of collaborations, communications and connection – in a synergetic way (Miron-Spektor et al., 2018) – between the poles of tensions to resolve them.

6.4.3 Ethical nuances in human-algorithm work interactions

The ethical dimensions of AI/algorithm technologies have been critically discussed in the literature (e.g., Kraemer et al., 2010; Amitai and Oren, 2017; Tsamados et al., 2022; Roberts et al., 2021). The adoption of algorithmic risk predictors for criminal justice work practices is received with mixed views in the literature. On the one hand, scholars such as Schwerzmann (2021) argue that the use of predictive tools in such institutions should be abolished as they are iteratively biased. On the other hand, there is empirical research that clearly suggests the integration of these tools positively contributes to the identification of human biases in the criminal justice's decision-making and minimises its risks (Kleinberg et al., 2018). The argument put forward by scholars such as Schwerzmann (2021) to completely abandon the algorithmic tools in criminal justice settings might sound polemic and based on cynicism. However, the literature is clear on the issue of algorithm bias, specifically in criminal justice organisations, regardless of whether an assessment is carried out fully or [semi]autonomously by algorithms (Hartmann and Wenzelburger, 2021). Risk of bias is plausible due to innate biases in statistical data and devaluation of human professionalism in favour of the algorithmic intelligence (McKay, 2020).

Findings illustrate that many data scientists have admitted that algorithmic predictions are far from ideal and may never be the best estimate to predict a risk, especially when it comes to sensitive and volatile tasks such as risk of offence. The data experts blame the lack of suitable inclusive data for developing risk predictors in their organisation. Therefore, it is possible that ethical issues, including biased decisions or discriminatory assessment, arise due to the unavailability of appropriate data. These ethical questions are akin to arguments by Tsamados et al. (2022): They argue and devise as an ethical framework that can help identifying and tackling biases in the data and algorithmic predictions. They highlight that having a fair, impartial and objective algorithm is tangled with the existence of non-discriminatory data, free from biases/stereotypes that affect variables in data such as race, gender, or age. However, other studies suggest that having bias-free algorithms can be challenging, as there is no simple way to filter bias from variables in the data (Mittelstadt et al., 2016). These ethical concerns were somewhat reflected in this research's findings. Many data scientists, frontline practitioners, and senior leaders

understand that there are limitations because of the biases in the data, particularly in demographics, which is a high risk of injustice in relation to risk assessment. This research, therefore, joins the theoretical conversations by Floridi and Taddeo (2016), Mittelstadt et al. (2016), and Tsamados et al. (2022), highlighting the importance of using appropriate and suitable data for algorithm development, and provides empirical evidence based on people's discourses on this ethical matter.

Another ethical dimension raised by some practitioners is the issue of biased feedback loop due to use of algorithms. It is a well-established argument in the AI/algorithm ethics literature that human biases may result in an abundance of biased data, which might then be used to design algorithms inflated by bias (Leicht-Deobald et al., 2019). Hence, any prediction, decision, or assessment made by the biased algorithms may, subsequently affect people's judgments, leading to more biased decisions (Mansoury et al., 2020). Some frontline practitioners specifically underline this as a pivotal ethical issue affecting risk assessment. Those employees are aware that their biases or prejudices in decision-making may add more biased data into the ARA tools. This is in line with the argument raised by Sun et al. (2020) in the sense that biased assessment from a human end-user may result in the production of biased data and eventually an algorithm. Subsequently, the biased algorithm generates predictions that might influence frontline officers' mindsets and professional judgements. Thereby, this research provides further empirical evidence on the ethical issue of 'biased feedback loop' (Sun et al., 2020). Furthermore, the identification of this ethical concern by some practitioners, signifies Foucauldian notions of self-reflexivity and self-awareness (Crane et al., 2008) of these actors in relation to their work experience with algorithms. Also, it indicates practitioners' activism to ensure the delivery of ethical conduct (Raffnsøe et al., 2019) despite the potential ethical flaws of algorithms. Hence, this research re-articulates and adds 'ethics of subjectivity' developed by Foucault (2020c) to existing AI/algorithms literature and emphasises the importance of human end-users' discourse of self-reflexivity in identification of ethical shortcomings of algorithms.

Notably, there are two important ethical aspects that are evident amongst the frontline practitioners rather than senior leaders or data analysts. First is the issue of categorisation and reduction of individuals, or 'datafication' of individuals (Murray and Flyverbom, 2020). Second is overreliance on algorithmic predictions and marginalisation of human intuitions and professional judgments. These two issues were unfolded by a number of practitioners as two prominent ethical dilemmas facing algorithmic criminal justice service. There are studies such as Tsamados et al. (2022) and Taddeo and Floridi (2018) that have conceptually – and succinctly – pointed out these two ethical questions in the AI/algorithm ethics literature. This research provides theoretical and further empirical support for the

discussions raised by Tsamados et al. (2022) regarding overreliance on algorithmic outputs as well as De Laat's (2019) argument on how algorithms can lead to decay of human agency. Current research in this area is based on viewpoints of academics with only limited empirical evidence to support the claims (e.g., Yu et al., 2018; Etzioni and Etzioni, 2017; Tsamados et al., 2022). Furthermore, much of the existing research aims to rectify ethical concerns via knowledge of computer/data science. More research is required in organisational contexts and work experiences of human end-users. As such, this study argues that practitioners should be at the centre of understanding the ethical nuances of algorithms, as they are the end-users whose working experiences shape the algorithmic work practices. As findings demonstrate, there were insightful statements from some practitioners that can expand the knowledge concerning algorithmic agential power (De Laat, 2019). In addition, this study argues that identifying the highlighted ethical issue has had a galvanising effect on some practitioners work experiences. As such, many practitioners have aimed to ensure their professional judgment is not totally influenced by the predictions of ARA tools (Introna, 2016).

Existing guidelines that predominantly contain principles of algorithmic ethics, and they include how the integration of those ethical principle may offer solutions to resolving ethical issues of algorithmic work (see: e.g., Tsamados et al. 2022; Taddeo and Floridi, 2018). The relevant ethics scholarship hitherto has scarcely incorporated discourses of those who directly interact with algorithms. This study has incorporated discourses from organisational actors; particular discursive patterns that Foucault identifies as enunciative or systems of power that can enable us to understand discontinuities, such as challenges, breaks, transformations, and practices (Foucault, 1972). These discourses can be materialised into something more concrete in algorithm ethics and constitute a different perspective from what is already known in the relevant scholarship. Both Foucault (1972) and Fairclough (1993) suggest contextualising unique discourses from the individuals that can shape the "social body." Following these conceptualisations, I argue that the known principles in algorithm ethics literature relatively overlook the perspectives of those who interact with algorithms. This study, however, brings to light the discursive patterns of the users of algorithms, including datafication and overreliance due to extensive utilisation of algorithms, and expands/modifies the existing AI/algorithm literature by emphasising the perspectives of human end-users around ethics.

Findings indicate that scientific power/knowledge featured in the art of governmentality (Seeck, 2011) aims to circumvent and assimilate other power/knowledge discourses in order to ensure the self-regulation of practitioners. As such, it seems palpable that scientific power/knowledge may continue to intensify (Hardy and Thomas, 2014) and

introduce more self-disciplinary technologies (e.g., advanced AI). However, irrespective of the dominance of this scientific power (Barratt, 2002; Foucault, 1977), other organisational actors are capable of exercising their own power/knowledge around the algorithmic practices. As such, this research advances AI/algorithm ethics literature by highlighting that ethical aspects are not solely about creating technologies that are less biased or discriminatory, but it includes the viewpoints and actions of those who interact with these technologies and refuse to be dominated by algorithmic agential power.

6.4.4 Signs of resistance against ARA commodities

The premise of workforce resistance against managerial orthodoxies has been widely discussed organisational studies (Van Dijk and Van Dick, 2009; Tucker, 1993; Dalgliesh, 2009), particularly in association with emerging technologies at work such as AI and algorithmic tools (Cameron and Rahman, 2022; De Vaujany et al., 2021; Bucher et al., 2021). Much of the discussion in this strand of literature focuses on employee resistances against managerial control and surveillance through AI/algorithms (e.g., Cameron and Rahman, 2022; Pignot, 2021; Kellogg et al., 2020). There is little known on how employees yield their resistance against other types of algorithmic work practices, for instance, when employees are situated within an intelligent work environment and need to work alongside algorithm technologies. As such, there are calls in the existing literature to further investigate how the workforce exercises its resistance against algorithmic commodities at work (Newlands, 2021; Jarrahi et al., 2021). In a similar manner to the existing studies, I illustrate that some frontline practitioners resist ARA tools, not in a sense to question the existence of them but to exercise their volition in outlining the ethical dilemmas associated with their use. In order to unpack this, I have invoked Foucauldian concepts of objectification and subjectification (Heller, 1996) and resistance (Dalgliesh, 2009), offering three important contributions to the literature on workforce resistance against AI/algorithms.

First, my findings shed empirical light on the underexplored notion of resistance against algorithmic work regimes (Newlands, 2021) in a criminal justice organisational setting where algorithms are used as predictors. The majority of studies in relation to resistance in organisational context remain conceptual, calling for more substantial evidence to scrutinise this issue (e.g., Alakavuklar and Alamgir, 2018; Raffnsøe et al., 2016). Furthermore, many empirical studies around employee resistance focus on particular workplace algorithms with a surveillant nature for employee monitoring or workforce management (e.g., Anteby and Chan, 2018; Pignot, 2021; De Vaujany et al., 2021). Although one might argue ethical questions associated with workforce algorithmic

surveillance may overlap with work-augmentative agents, there is only limited academic insight on how organisational actors resist these intelligent counterparts in their work interactions. By exploring the context of the criminal justice service, this research discovers that resistance against this particular algorithmic work regime is about ethical work practices (i.e., risk assessment) and fills this lacuna empirically.

Second, this study brings to light frontline practitioners' discourses that can inform algorithm ethics. As suggested in the findings, forms of resistance have shifted from collectivist to more individualistic types in a way that frontline employees no longer shout their concerns or protest ARA tools explicitly. Many practitioners are now more inclined to show their dissatisfaction or concerns by criticising the tools and outlining the ethical ramifications associated with them. As such, this study unfolds actions implemented by those practitioners to ensure that they do not let ARA steer their professional judgment. Practitioners' perspectives on how ARA reduces individuals to numbers or conducts such as *professional overriding* the ARA predictions are in line with the Alakavuklar and Alamgir (2018) study: They propose workforce resistance against managerial power/knowledge is embodied in their practical discourses and formations in their everyday working lives. Foucault (2020c) argues that subjects who have been objectified by the exercise of power/knowledge may question the dominant discourse through their self-awareness. They become aware of the disciplinary mechanisms of the power and act to constitute their own sense of individuality or freedom (Gollmitzer, 2023) within power mechanisms. In other words, people's resistance is crystallised through their discursive practices – actions – with the aim to transform social milieux (Clegg et al., 2006) and resist being objectified (Heller, 1996). In the context of algorithmic criminal justice, some frontline practitioners have acted as vanguards in exercising resistance against scientific power/knowledge dominance. Those who acted upon themselves – through 'subjectification' (Foucault, 1980) – and added novel ethical discourses in relation to ARA tools. This study illustrates that although algorithmic technologies act as new forms of disciplinary power/knowledge mechanisms to regulate – or rather standardise – traditional risk assessment practices, many frontline practitioners were able to challenge a dominant discourse and recognise the ethical issues of algorithmic predictors.

And finally, this study uncovers the genesis of ethical subjects (Crane et al., 2008) in the context of algorithmic criminal justice organisation. As it was explained above, the highlighted subtle acts of discursive resistance have resulted in novel creative forms of ethics around ARA tools. As such, the practitioners' journey towards becoming self-aware and self-reflexive has led to the constitution of ethical subjects, according to Foucault's (1988) later work on *Ethics* and *Care for the Self*. Scholars who have problematised this

dimension of Foucault argue that his ethics project is particularly relevant to notions such as individual's liberty and emancipation (e.g., White, 2014). Organisational scholars such as Crane et al. (2008) and Skinner (2013) have made calls to encourage scholars to investigate the ethical dilemmas of organisation through the lens of Foucauldian *Ethics* and *Care for the Self*. In line with Foucault's concept of ethics, White (2014) highlights that ethical subjects are not those who believe in morality by law/code, but they are those who detach themselves from domination of rules. Following on this problematisation, I argue that as many organisational stakeholders, especially frontline practitioners, resist the domination of scientific power relations and unpack novel ethical discourses, they also form themselves as Foucauldian *ethical subjects*. Being moral or ethical, for Foucault, does not *per se* encompass those organisational stakeholders who conform to disciplinary power, follow the enacted ethical toolkits, and alienate themselves from self-consciousness. Foucauldian ethical subjects (Raffnsøe et al., 2019) – in the context of ARA criminal justice – are those who have shown their resistance by expressing their concerns around ARA practice, criticizing the existing ethical principles through their own discourses. This study joins Weiskopf and Hansen's (2023) conversation and argues that the 'space for ethical conduct' exists as organisational actors question, talk, and express their concerns around algorithms as well as subtle discursive resistance, which is a feature of this study.

6.5 Limitations and Avenues for Future Research

This qualitative study offers an empirical exploration of the ethical discourses of the key organisational stakeholders around the deployment of algorithms in a European criminal justice service. As in any research study, important limitations are associated with my claims, which could offer potential avenues for future research.

First, I focus on the analysis of interviews and documentation as the main sources of data, whilst there were employee surveys in which frontline practitioners have already reflected on their experiences around the utilisation of ARA tools. Because of the lack of access to the highlighted survey data, it is difficult to confirm whether there are viewpoints that could further inform ARA ethics. The lack of that data also limits this study's ability to verify whether and how the data scientists of organisation tackle the ethical questions raised by the frontline employees. Having access to the survey data would be undoubtedly priceless to further understand the intensification and dominance of power/knowledge (Hardy and Thomas, 2014) through scientific practices. Furthermore, as it is argued through Foucauldian theory, the resistances against the disciplinary mechanisms of power could be encapsulated in the discourses of those individuals who are objectified by power (Bergström

and Knights, 2006). I speculate that the survey data might entail particular viewpoints, concerns, and laments around ARA practice that could further support my interpretations on the notion of resistance against the ARA work regime. Of course, the absence of the survey data does not diminish this study's contributions since it focuses on uncovering the ethical algorithms' ethical discourses from a Foucauldian theoretical lens. The data collected through interviews was limited due to time constraints and the level of access granted to the researcher. However, it was continued up to the point of data saturation which could answer the research questions as it is argued by Bryman (2016). Future research could use other techniques of data collection, such as 'ethnography' (Brewer, 2000), that could guarantee data triangulation and further ensure the credibility of qualitative research (Lincoln and Guba, 1985).

Second, the choice of a qualitative-exploratory strategy for this research was purely due to its main aim to understand dimensions of algorithm ethics through the eyes of key organisational stakeholders as well as the probing nature of the research questions. Therefore, this study was not able to measure the extent to which the perception of ethics can affect particular employee behaviours or other organisational outcomes. For instance, through the analysis, I have uncovered that there are certain ethical concerns associated with the use of ARA technologies. In addition, the reactions and attitudes of practitioner end-users towards ARA tools were unearthed. But there is only little insight in my analysis on the associations and impacts of practitioners' positive/negative perceptions on the acceptance of ARA tools. Similarly, little is known around the (in)direct impacts of various perceptions on the overall performance of organisational actors. Future research may adopt a deductive approach and design quantitative studies that could better analyse the influence of perception of ethics on different aspects of employee behaviour. I encourage future research to design and test models that can test how ethical dimensions of algorithms – highlighted in this research – could negatively/positively affect, be affected or moderate different employee attitudes and/or other organisational phenomena.

Third, my sampling strategy and the level of regional access did not allow us to have the perfect representation of the organisational stakeholders across the nation. Criminal justice organisations are part of civil service operating at national level. The case study organisation in this research consists of 12 regional divisions that employ over thousands of individuals with various roles and responsibilities. And each of these individuals has their own unique views, perspectives, and discourses around the ethical aspects of ARA tools. As such, I speculate that the discourses from the frontline practitioners could diverge from region to region depending on the demographics and specifications of regional communities (e.g., the diversity of the community, the crime rate, etc.). These factors could potentially

impact the human professionals' interactions with ARA tools and shape various discourses around the ethics. Future research could consider the prevalence of these elements and explore whether the ethical discourses are influenced by the demographics of regions and communities.

Fourth, invoking Foucauldian theories marginally fails to inform how decentred agency of organisational stakeholders can initiate a change in the algorithmically enabled criminal justice system. In this criticism, I follow Caldwell's (2007) argument that although the Foucauldian notion of agency situates subjects in the form of self-reflexivity and transformation of the self, it also puts them in a state of flux and self-doubt. Indeed, Foucauldian concepts of subjectification and resistance are instrumental in exploring how subjects seek to change organisation and societies within the disciplinary power/knowledge discourses (Skinner, 2013). But it fails to synthesise a route that could lead to credible change or "making a difference" (Caldwell, 2007). Thus, I argue that although incorporating Foucault's standpoints in resistance and activism has not limited this research in understanding the ethical discourses around the utilisation algorithmic predictors. Yet, more is needed to understand how intentional agency and resistance against algorithmic work orthodoxies can be mediated through practices to make actual and positive moral-political actions and changes. Perhaps future research can build theoretical frameworks that will go beyond discursive agency, subjectification, and resistance, and examine how power/knowledge-based self-formations against algorithms can constitute tangible ethical changes in 4.0 intelligent organisations.

And finally, as the 4.0 technologies are becoming more ubiquitous in the context of work, it is only sensible to anticipate that the workplace could be reconfigured (Cameron and Rahman, 2022) and novel ethical dilemmas will emerge (Newlands, 2021). Therefore, there is always a need to re-examine the existing organisational theories and update them according to paradigm shifts of organisations (Barley et al., 2017). This research has explored the discourses that circulate amongst the key organisational stakeholders with regards to ethical questions associated with the use of algorithms at work. In doing so, I also unearthed particular dominating power/knowledge discourses by data science as well as frontline practitioners acts of subjectification and resistance against them. Yet, as the context of work transforms due to the rise of AI and algorithms, I assume that novel ethical questions will emerge, and hence, people will react differently to these emerging challenges. And as these challenges continue to grow, the organisational scholarship requires innovative data collection methods and theoretical frameworks to explore the emerging discourses.

6.6 Conclusion

This chapter has discussed the findings of this study in connection to the relevant literature. Through a Foucauldian theoretical lens, this research has spoken to and further extended organisational ethics literature by unearthing the art of governmentality and utilisation of algorithmic technologies. In that regard, this research has argued that ethical work practice, from the perspective of governmentality, is through the facilitation of algorithmic work environments that proliferate consistent bias-free decision-making. The art of governmentality supports and glorifies the power/knowledge of data science whilst marginalises other power/knowledge relations, including those of frontline practitioners. With that in mind, this research has applied and re-articulated the Foucauldian premise of 'governmentality' to show that the dominance of scientific power/knowledge is seen as the ethical solution to the risk of biases in human-based decision-making. This is a crucial point since, from an organisational ethics literature perspective, the government aims to make practitioners self-regulated by ensuring their conformity to the utilisation of algorithms, marginalising their voices and power/knowledges.

This study has also made further contributions to the literature on ethics of AI/algorithms as well as human-algorithm work collaborations. Adopting Foucault's later works on subjectification, self-reflexivity and resistance, this thesis argues that many organisational actors have been able to analyse the disciplinary mechanisms of governmentality and algorithmic work. In that regard, the actors, in particular, some frontline practitioners, have shed light on different aspects of using ARA tools, including the flaws, limitations, and ethical nuances, such as how algorithmic prediction could misdirect or cloud their judgments. This study has highlighted the "aesthetics" (Raffnsøe et al., 2019) of practitioners' conduct in their interactions with ARA, outlining the extent to which their subtle contestation and resistances are formed against algorithmic practices. The implicit and subtle resistance is specifically coordinated to minimise the impact of governmentality via algorithms. This study, therefore, moves a step further from the dominance of computer data science in existing algorithm ethics literature and advances it by including the human factors, how human end-users exercise their agency to become 'ethical individuals.'

In the next chapter, conclusions, I summarise this thesis's main findings and also outline a number of managerial implications derived from the data of this research.

CHAPTER 7: Conclusion

Algorithmic work practice has proliferated in many organisational settings, indicating that AI/algorithms are no longer a fad. It has massively transformed the way the organisational actors interact with their tasks and work processes. As the premise of 'ethics' in algorithmic work expands in scholarship, it remains rather new and theoretical within organisational contexts, in need of further scrutiny. Discourse of ethics in organisational scholarship was highlighted from two angles in the existing research: firstly, from a technological/scientific point of view and the way the technology itself might cause ethical issues (Tsamados et al., 2022; Mittelstadt et al., 2016), which predominantly entails publications from the computer science domain. And secondly, from a critical organisational literature lens, which aims to understand the role, agency, and perspectives of humans in algorithmic work practices. In this thesis, I used empirical evidence from a series of qualitative interviews from the actors' perspectives of a criminal justice organisation to illustrate that ethics are existent and ubiquitous. However, this study indicated that it is perceived differently by organisational actors. Whilst for the senior managers and data scientists, ethics is crystallised in the use of algorithms, the findings illustrate that at least for some human frontline practitioners, algorithms are not innocuous technologies. Furthermore, for the frontline practitioners, algorithms are seen as augmentation tools that would help them to be more consistent and transparent in their work practices. Ethics, for frontline practitioners, is intertwined in their practices or delivery of a fair/transparent conduct through the use of algorithms.

By drawing on Foucauldian concepts, this research advances the theories of AI/algorithm ethics from the peculiar angle of organisational stakeholders, different from existing studies. Previous research on AI/algorithms ethics has cast light on the issues with the use of algorithms that might negatively impact people, including in-built biases or threats to human autonomy. This study, however, not only provided empirical and contextualised evidence for the mentioned claims in the literature but also went further and uncovered particular power/knowledge rationalities that implement algorithms as vessels for ethical conduct (i.e., through the art of governmentality and advocating discourses of data science expertise). With this in mind, I conclude this thesis with a few suggestions that contribute to organisational and managerial practice.

First, this study helps the data scientists to understand that the issues, such as biased discriminatory risk assessments, may emerge due to the unavailability of appropriate data. As the analysis suggests, the data that is incorporated in the design and development of ARA technologies is not particularly collected for the purpose of AI/algorithms, but for the administration of individuals. Although this limitation *per se* has not affected the design

process, it is deemed problematic by the organisational players and has engendered several ethical challenges. As such, many data scientists and frontline practitioners believe that ARA tools are not producing impartial, objective predictions. Therefore, I suggest that the data that is used and fuelled to ARA should be collected solely for the purpose of algorithm design. Furthermore, this collected data should be evaluated and assessed for any ethical ramifications, such as systemic bias and stereotypes, as suggested by Floridi et al. (2018). By utilising proper datasets in ARA technologies, the criminal justice service can ensure the transparency, accountability (Ananny and Crawford, 2018) and trustworthiness (Tsamados et al., 2022) of their ARA practices.

Second, as the findings of this study indicate, the processes of design, development, and utilisation of algorithmic tools were predominantly steered by the expertise of data scientists and engineers and subsequently endorsed by the politicians and senior management. It means that voices and inputs of human end-users (i.e., the frontline practitioners) were excluded from the implementation processes, which then made the operationalisation of the algorithm difficult within this organisation. To explain more, the adoption of ARA technologies is seen as a partial success because it is received with mixed feelings by some of the frontline practitioners. Therefore, I recommend that organisational leaders, together with data science teams, develop better communication channels and interventions that foster the understanding, cooperation, and collaboration between divisions, teams, and other stakeholders (Chowdhury et al., 2023). By doing so, the operationalisation and mobilisation of the algorithm will better take place and maximise the effectiveness and acceptance of the intelligent tools. Concurring with the idea of dignity in organisational ethics (Phillips and Margolis, 1999), I argue that including frontline practitioners in the ARA tools' design, not only would it benefit the ethical dimension of those tools, but it would also strengthen the acceptance of technologies amongst the users (Stamate et al., 2021).

Third, the study reveals that this criminal justice setting lacks an ecosystem that supports interdisciplinary collaboration, or rather, a mechanism for knowledge sharing for the continuous development of competencies, skills, and knowledge amongst the end-user practitioners in the context of ARA tools. In that regard, the findings outline that the practitioner end-users have received only limited training on ARA tools and their internal processing. Such trainings are delivered not by data science teams but by senior practitioners who took the role of instructors. Further training was provided only via documentation in the form of user guidelines, containing dos and don'ts. The lack of knowledge sharing and interdepartmental communication has left some practitioners rather unaware or reluctant to seize all new opportunities created by ARA practices. In that regard,

I recommend that managerial strategies enact a more viable knowledge-sharing ecosystem (Lauring and Selmer, 2012) responsible for evolving the transfer of knowledge to the frontline practitioners. This strategy can help employees to better embrace algorithms and their opportunities and better respond to the volatility of their work practice.

Fourth, the study unfolded that some end-user practitioners are sceptical of their work interactions with algorithms and expressed relatively negative viewpoints around algorithmic predictions. As such, they have invoked behaviours that show their contestation and subtle resistance against ARA, including disregard or exclusion of algorithms from their risk assessments or overriding the generated outputs. This has resulted in mistrust and alienation of the practitioners (Dietvorst et al., 2015) in algorithmic processes. To rectify this, I suggest that senior managers and data scientists better outline and communicate their ARA strategies with the frontline practitioners. Better communication of ARA strategies means clearly and transparently explaining the purpose of using ARA technologies, including the benefits and limitations, the impacts on the day-to-day working tasks, the roles and responsibilities of the practitioners, and the expectation from them. This research highlighted some of the issues and chasms in relation to inter-departmental communications in the organisations. Thus, moving forward, the organisation can benefit from a transparent multilevel algorithm work strategy that cements effective implementation of these tools.

And finally, according to a report published by CIPD (2018), there is a growing fear amongst the general public that the introduction of intelligent technologies means proliferation of managerial control, monitoring, and surveillance. As such, the consequence of utilising algorithms for cognitive augmentation of humans is depicted as the decay of human autonomy and professionalism. Although the algorithms used in the context of this research were not designed to monitor and/or control the working lives of human practitioners, findings indicate that a few practitioners had had concerns around the future landscape of AI/algorithms. These concerns, however minimal, illustrate the people's uncertainties of a future where human autonomy and agency are replaced by machine intelligence. Indeed, as research suggests, these concerns are somewhat valid, such as the issue of overreliance on AI and algorithmic tools. However, it is within the responsibility of senior leaders and HR professionals to consider these concerns and give reassurance that algorithms bring enrichment to the job. It is crucial for the human end-users to be aware that algorithms are not taking control but bringing more flexibility and productivity to the work processes.

To sum up, this thesis adopted a Foucauldian perspective to better understand the ethical dimension of utilising algorithms in a criminal justice organisational setting. This study

illustrated that whilst for senior leaders and the data team, ethics is aligned with the use of algorithms, frontline actors in this organisation have acted upon to ensure their service delivery is based on transparency, fairness, and impartiality. This study shows that space for ethics exists not only in the implementation of algorithms but also in the perspectives, discourses, and actions of organisational actors. As such, this research suggests that the discourse of ethics in algorithmic work environments is shaped by organisational actors through their agency and self-reflexivity.

REFERENCES

Abara IOC and Singh S (1993) Ethics and biases in technology adoption: The small-firm argument. *Technological Forecasting and Social Change* 43(3-4): 289-300.

Abuhusain M (2020) The role of artificial intelligence and big data on loan decisions. *Accounting* 6(7): 1291-1296.

Adadi A and Berrada M (2018) Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), in *IEEE Access*, vol. 6, pp. 52138-52160, 2018, doi: 10.1109/ACCESS.2018.2870052. keywords: (Conferences; Machine learning; Market research; Prediction algorithms; Machine learning algorithms; Biological system modelling; Explainable artificial intelligence; interpretable machine learning; black-box models),

Agar M (1996) *The Professional Stranger: An Informal Introduction to Ethnography.* Emerald Group Publishing Limited.

Ahonen P, Tienari J, Meriläinen S, et al. (2014) Hidden contexts and invisible power relations: A Foucauldian reading of diversity research. *Human Relations* 67(3): 263-286.

Ajunwa I (2020) The “black box” at work. *Big Data & Society* 7(2).

Alaimo C and Kallinikos J (2021) Managing by Data: Algorithmic Categories and Organizing. *Organization Studies* 42(9): 1385-1407.

Alakavuklar ON and Alamgir F (2018) Ethics of resistance in organisations: A conceptual proposal. *Journal of Business Ethics* 149(1): 31-43.

Allen M (2019) Chapter 3 - Artificial intelligence (AI). In: Allen M (ed) *The Chief Security Officer's Handbook*. Academic Press, pp.35-65.

Allen RT and Choudhury P (2022) Algorithm-Augmented Work and Domain Experience: The Countervailing Forces of Ability and Aversion. *Organization Science* 33(1): 149-169.

Alsheibani S, Cheung Y and Messom C (2018) Artificial Intelligence Adoption: AI-readiness at Firm-Level. *Artificial Intelligence* 6: 26-2018.

Althusser L (1969) *For Marx.* Allen Lane.

Alvesson M and Karreman D (2000) Varieties of Discourse: On the Study of Organizations through Discourse Analysis. *Human Relations* 53(9): 1125-1149.

Amicelle A (2022) Big data surveillance across fields: Algorithmic governance for policing & regulation. *Big Data and Society* 9(2).

Amitai E and Oren E (2017) Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics* 21(4): 403-418.

Ananny M and Crawford K (2018) Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media and Society* 20(3): 973-989.

Anteby M and Chan CK (2018) A Self-Fulfilling Cycle of Coercive Surveillance: Workers' Invisibility Practices and Managerial Justification. *Organization Science* 29(2): 247-263.

Aoki N (2020) An experimental study of public trust in AI chatbots in the public sector. *Government Information Quarterly* 37(4): 101490.

Arribas-Ayllon M and Walkerdine V (2017) Foucauldian Discourse Analysis (Second Edition). *The Sage handbook of qualitative research in psychology*. pp.110-123.

Atkinson P, Delamont S and Coffey A (2004) *Key Themes in Qualitative Research: Continuities and Changes.* AltaMira Press.

Au YA, AU YA and KAUFFMAN RJ (2003) What do you know? Rational expectations in information technology adoption and investment. *Journal of Management Information Systems* 20(2): 49-76.

Bader V and Kaiser S (2019) Algorithmic decision-making? The user interface and its role for human involvement in decisions supported by artificial intelligence: The Interdisciplinary Journal of Organization, Theory and Society. *Organization* 26(5): 655-672.

Bakir V (2015) "Veillant Panoptic Assemblage": Mutual Watching and Resistance to Mass Surveillance after Snowden. *Media and Communication* 3(3): 12-25.

Barley SR, Bechky BA and Milliken FJ (2017) The Changing Nature of Work: Careers, Identities, and Work Lives in the 21st Century. *Academy of Management Discoveries* 3(2): 111-115.

Barratt E (2002) Foucault, Foucauldianism and human resource management. *Personnel Review* 31(2): 189-204.

Barratt E (2003) Foucault, HRM and the Ethos of the Critical Management Scholar. *Journal of Management Studies* 40(5): 1069-1087.

Barratt E (2008) The later Foucault in organization and management studies. *Human Relations* 61(4): 515-537.

Barry L (2019) The rationality of the digital governmentality. *Journal for Cultural Research* 23(4): 365-380.

Baum SD (2020) Social choice ethics in artificial intelligence. *AI & SOCIETY: Journal of Knowledge, Culture and Communication* 35(1): 165.

Baxter G and Hainey T (2024) Using immersive technologies to enhance the student learning experience. *Interactive Technology and Smart Education* 21(3): 403-425.

Behrent M C (2013). Foucault and Technology. *History and Technology*, 29(1), 54–104. https://doi.org/10.1080/07341512.2013.780351

Bell E, Bryman A and Harley B (2018) *Business Research Methods.* Oxford University Press.

Bergen, J.P., Verbeek, PP (2021) To-Do Is to Be: Foucault, Levinas, and Technologically Mediated Subjectivation. *Philosophy and Technology*. 34, 325–348. https://doi.org/10.1007/s13347-019-00390-7

Bergström O and Knights D (2006) Organizational discourse and subjectivity: Subjectification during processes of recruitment. *Human Relations* 59(3): 351-377.

Bergström O and Knights D (2016) Organizational discourse and subjectivity. *Human Relations* 59(3): 351-377.

Bernauer J and Rasmussen D (1988) *The Final Foucault.* MIT Press.

Bevir, M. (1999). Foucault, Power, and Institutions. *Political Studies*, 47(2), 345-359. https://doi.org/10.1111/1467-9248.00204

Bigman YE, Wilson D, Arnestad MN, et al. (2022) Algorithmic discrimination causes less moral outrage than human discrimination. *J Exp Psychol Gen*. Epub ahead of print 2022/06/28. DOI: 10.1037/xge0001250.

Blaikie N (2009) *Designing Social Research: The Logic of Anticipation.* Wiley.

Boden MA (2018) *Artificial Intelligence: A Very Short Introduction.* OUP Oxford.

Braun V and Clarke V (2019) Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11(4): 589-597.

Braun V, Clarke V and Hayfield N (2022) 'A starting point for your journey, not a map': Nikki Hayfield in conversation with Virginia Braun and Victoria Clarke about thematic analysis. *Qualitative Research in Psychology* 19(2): 424-445.

Brewer J (2000) *Ethnography.* McGraw-Hill Education.

Bryman A (2003) *Quantity and Quality in Social Research.* Taylor & Francis.

Bryman A (2016) *Social research methods.* Oxford University Press.

Bryman A and Burgess B (2002) *Analyzing Qualitative Data.* Taylor & Francis.

Bucher EL, Schou PK and Waldkirch M (2021) Pacifying the algorithm – Anticipatory compliance in the face of algorithmic management in the gig economy. *Organization* 28(1): 44-67.

Budhwar P, Chowdhury S, Wood G, et al. (2023) Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT. *Human Resource Management Journal* 33(3): 606-659.

Buhalis D and Leung R (2018) Smart hospitality—Interconnectivity and interoperability towards an ecosystem. *International Journal of Hospitality Management* 71: 41-50.

Burrell G (1998) Linearity, Control and Death. In: *Discourse and Organization*. David Grant, Tom Keenoy and Cliff Oswick Editors, London: SAGE Publications Ltd. pp. 135-151 Available at: https://doi.org/10.4135/9781446280270.

Burrell G and Morgan G (2017) *Sociological Paradigms and Organisational Analysis: Elements of the Sociology of Corporate Life.* Taylor & Francis.

Burton JW, Stein MK and Jensen TB (2020) A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33(2): 220-239.

Busch C, Crawshaw J, Guillaume Y, et al. (2024) Ethics-Related Mentoring: A Scale Development and Test of its Role in Promoting Protégé Ethical Behaviour. *British Journal of Management* 35(1): 210-227.

Butt AH, Ahmad H, Goraya MAS, et al. (2021) Let's play: Me and my AI-powered avatar as one team. *Psychology and Marketing* 38(6): 1014-1025.

Caldwell C, Truong DX, Linh PT, et al. (2010) Strategic Human Resource Management as Ethical Stewardship. *Journal of Business Ethics* 98(1): 171-182.

Caldwell R (2007) Agency and Change: Re-evaluating Foucault's Legacy. *Organization* 14(6): 769-791.

Cameron LD and Rahman H (2022) Expanding the locus of resistance: Understanding the co-constitution of control and resistance in the gig economy. *Organization Science* 33(1): 38-58.

Campbell C (2009) Distinguishing the power of agency from agentic power: A note on Weber and the “black box” of personal agency. *Sociological theory* 27(4): 407-418.

Campbell C, Sands S, Ferraro C, et al. (2020) From data to action: How marketers can leverage AI. *Business Horizons* 63(2): 227-243.

Carmine S and De Marchi V (2023) Reviewing Paradox Theory in Corporate Sustainability Toward a Systems Perspective. *Journal of Business Ethics* 184(1): 139-158.

Chan SC and Ngai EW (2007) A qualitative study of information technology adoption: how ten organizations adopted Web-based training. *Information Systems Journal* 17(3): 289-315.

Charlwood A and Guenole N (2022) Can HR adapt to the paradoxes of artificial intelligence? *Human Resource Management Journal* 32(4): 729-742.

Charmaz K (2014) *Constructing Grounded Theory.* SAGE Publications.

Chatterjee S, Chaudhuri R, Vrontis D, et al. (2023) Adoption of blockchain technology in organizations: from morality, ethics and sustainability perspectives. *Journal of Information, Communication and Ethics in Society* ahead-of-print(ahead-of-print).

Chatterjee S, Kar AK and Gupta MP (2018) Success of IoT in Smart Cities of India: An empirical analysis. *Government Information Quarterly* 35(3): 349-361.

Chatterjee S, Rana NP, Dwivedi YK, et al. (2021) Understanding AI adoption in manufacturing and production firms using an integrated TAM-TOE model. *Technological Forecasting and Social Change* 170.

Chen H, Li L and Chen Y (2021) Explore success factors that impact artificial intelligence adoption on telecom industry in China. *Journal of Management Analytics* 8(1): 36-68.

Chornous GO and Gura VL (2020) Integration of information systems for predictive workforce analytics: Models, synergy, security of entrepreneurship. *European Journal of Sustainable Development* 9(1): 83-83.

Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5(2): 153-163.

Chowdhury S, Dey P, Joel-Edgar S, et al. (2023) Unlocking the value of artificial intelligence in human resource management through AI capability framework. *Human Resource Management Review* 33(1): 100899.

Chye Koh H and Boo EfHY (2004) Organisational ethics and employee satisfaction and commitment. *Management Decision* 42(5): 677-693.

CIPD (2018) *People and Machines: From hype to Reality.* [pdf] CIPD – Chartered Institute for Personnel Development –. Available at: https://www.cipd.org/globalassets/media/zzz-misc---to-check/11people-and-machines-report-1_tcm18-56970.pdf [Accessed 10 April 2021].

Clark T, Foster L, Bryman A, et al. (2021) *Bryman's Social Research Methods.* Oxford University Press.

Clegg S (1994) Weber and Foucault: Social theory for the study of organizations. *Organization* 1(1): 149-178.

Clegg S (1998) Foucault, power and organizations.

Clegg S, Kornberger M and Rhodes C (2007) Business ethics as practice. *British Journal of Management* 18(2): 107-122.

Clegg SR (1989) *Frameworks of Power.* SAGE Publications.

Clegg SR, Courpasson D and Phillips N (2006a) *Power and Organizations.* SAGE Publications.

Clegg SR, Hardy C, Lawrence T, et al. (2006b) *The SAGE Handbook of Organization Studies.* SAGE Publications.

Clegg, S., and D. Dunkerley (1980) *Organization, class and control.* London: Routledge and Kegan Paul.

Cludts S (1999) Organisation Theory and the Ethics of Participation. *Journal of Business Ethics* 21(2): 157-171.

Cohen L, Manion L and Morrison K (2013) *Research Methods in Education.* Taylor & Francis.

Coombs C (2020) Will COVID-19 be the tipping point for the Intelligent Automation of work? A review of the debate and implications for research. *International Journal of Information Management* 55: 102182.

Cooper R (2020) Pastoral power and algorithmic governmentality. *Theory, Culture & Society* 37(1): 29-52.

Crane A, Knights D and Starkey K (2008) The conditions of our freedom: Foucault, organization, and ethics. *Business Ethics Quarterly* 18(3): 299-320.

Crawshaw JR (2006) Justice source and justice content: evaluating the fairness of organisational career management practices. *Human Resource Management Journal* 16(1): 98-120.

Creswell JW and Poth CN (2016) *Qualitative Inquiry and Research Design: Choosing Among Five Approaches.* SAGE Publications.

Crotty M (1998) *The Foundations of Social Research: Meaning and Perspective in the Research Process.* SAGE Publications.

Cunneen C (2006) Racism, discrimination and the over-representation of Indigenous people in the criminal justice system: Some conceptual and explanatory issues. *Current Issues in Criminal Justice* 17(3): 329-346.

Curchod C, Patriotta G, Cohen L, et al. (2020) Working for an algorithm: Power asymmetries and agency in online work settings. *Administrative Science Quarterly* 65(3): 644-676.

Dalgliesh B (2009) Foucault and creative resistance in organisations. *Society and Business Review* 4(1): 45-57.

Daugherty PR and Wilson HJ (2018) *Human + Machine: Reimagining Work in the Age of AI.* Harvard Business Review Press.

Davis AJ (1996) Benign Neglect of Racism in the Criminal Justice System. *Michigan Law Review* 94(6): 1660-1686.

Davis FD (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*. 319-340.

De Cremer D and McGuire J (2022) Human–Algorithm Collaboration Works Best if Humans Lead (Because it is Fair!). *Social Justice Research* 35(1): 33-55.

De Laat M, Joksimovic S and Ifenthaler D (2020) Artificial intelligence, real-time feedback and workplace learning analytics to support in situ complex problem-solving: a commentary. *International Journal of Information and Learning Technology* 37(5): 267-277.

De Laat PB (2018) Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability? *Philosophy and Technology* 31(4): 525-541.

De Laat PB (2019) The disciplinary power of predictive algorithms: a Foucauldian perspective. *Ethics and Information Technology* 21(4): 319-329.

De Vaujany F-X, Leclercq-Vandelannoitte A, Munro I, et al. (2021) Control and Surveillance in Work Practice: Cultivating Paradox in 'New' Modes of Organizing. *Organization Studies* 42(5): 675-695.

Deleuze G (2006) *Foucault.* Bloomsbury Academic.

Denzin NK and Lincoln YS (2011) *The SAGE handbook of qualitative research.* SAGE.

Derrida J (1978) *Writing and Difference.* University of Chicago Press.

Desouza KC, Dawson GS and Chenok D (2020) Designing, developing, and deploying artificial intelligence systems: Lessons from and for the public sector. *Business Horizons* 63(2): 205-213.

Diakopoulos N (2015) Algorithmic Accountability: Journalistic investigation of computational power structures. *Digital Journalism* 3(3): 398-415.

Diakopoulos N and Koliska M (2017) Algorithmic Transparency in the News Media. *Digital Journalism* 5(7): 809-828.

Dietvorst BJ, Simmons JP and Massey C (2015) Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1): 114.

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

Dreyfus HL and Rabinow P (1983) *Michel Foucault: Beyond Structuralism and Hermeneutics.* University of Chicago Press.

Du Plessis EM (2020) Speaking truth through power: Conceptualizing internal whistleblowing hotlines with Foucault's dispositive. *Organization* 29(4): 544-576.

Duan Y, Edwards JS and Dwivedi YK (2019) Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. *International Journal of Information Management* 48: 63-71.

Durán JM and Jongsma KR (2021) Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* 47(5): 329-335.

Dwivedi YK, Hughes L, Baabdullah AM, et al. (2022) Metaverse beyond the hype: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management* 66.

Dwivedi YK, Hughes L, Ismagilova E, et al. (2021) Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management* 57.

Edmondson AC and Mcmanus SE (2007) Methodological fit in management field research. *Academy of Management Review* 32(4): 1246-1264.

Eglash R, Robert L, Bennett A, et al. (2020) Automation for the artisanal economy: enhancing the economic and environmental sustainability of crafting professions with human-machine collaboration. *Ai & Society* 35(3): 595-609.

Elias S (2008) Fifty years of influence in the workplace: The evolution of the French and Raven power taxonomy. *Journal of Management History* 14(3): 267-283.

Etter M and Albu OB (2021) Activists in the dark: Social media algorithms and collective action in two social movement organizations. *Organization* 28(1): 68-91.

Etzioni A and Etzioni O (2017) Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics* 21(4): 403-418.

Fairclough N (1993) *Discourse and Social Change.* Wiley.

Fairclough N (2003) *Analysing Discourse: Textual Analysis for Social Research.* Routledge.

Fairclough N (2013) *Critical Discourse Analysis: The Critical Study of Language.* Taylor & Francis.

Falwadiya H and Dhingra S (2022) Blockchain technology adoption in government organizations: a systematic literature review. *Journal of Global Operations and Strategic Sourcing* 15(3): 473-501.

Farrokhi A, Shirazi F, Hajli N, et al. (2020) Using artificial intelligence to detect crisis related to events: Decision making in B2B by artificial intelligence. *Industrial Marketing Management* 91: 257-273.

Flick U (2013) The SAGE Handbook of Qualitative Data Analysis. London: SAGE Publications Ltd.

Floridi L (2018) Soft ethics and the governance of the digital. *Philosophy & Technology* 31: 1-8.

Floridi L (2019) Establishing the rules for building trustworthy AI. *Nature Machine Intelligence* 1(6): 261-262.

Floridi L and Sanders JW (2002) Mapping the foundationalist debate in computer ethics. *Ethics and Information Technology* 4(1): 1-9.

Floridi L and Taddeo M (2016) Introduction: What is data ethics? *Philosophical Transactions: Mathematical, Physical and Engineering Sciences* 374(2083): 1-5.

Floridi L, Cowls J, Beltrametti M, et al. (2018) AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines: Journal for Artificial Intelligence, Philosophy and Cognitive Science*. DOI: 10.1007/s11023-018-9482-5. 1.

Flyvbjerg B (1998) Habermas and Foucault: thinkers for civil society? *British Journal of Sociology*. 210-233.

Ford J and Harding N (2003) Invoking Satan or the Ethics of the Employment Contract. *Journal of Management Studies* 40(5): 1131-1150.

Foucault M (1972) *The Archaeology of Knowledge: And the Discourse on Language.* Knopf Doubleday Publishing Group.

Foucault M (1977) *Discipline and Punish: The Birth of the Prison.* Vintage Books.

Foucault M (1980) *Power/knowledge: Selected Interviews and Other Writings, 1972-1977.* Vintage Books.

Foucault M (1982) The Subject and Power. *Critical inquiry* 8(4): 777-795.

Foucault M (1983). The subject and power. In H. Dreyfus & P. Rabinow (Eds.), *Michel Foucault: Beyond structuralism and hermeneutics* (2nd ed., pp. 208–226). Chicago: University of Chicago Press.

Foucault M (1988) *Politics, philosophy, culture: Interviews and other writings, 1977-1984.* New York, NY, US: Routledge, Chapman & Hall.

Foucault M (1988) *The History of Sexuality, Vol. 3: The Care of the Self.* Knopf Doubleday Publishing Group.

Foucault M (1990) *The History of Sexuality: An Introduction.* Knopf Doubleday Publishing Group.

Foucault M (2018) *The Order of Things.* Taylor & Francis.

Foucault M (2019) *The History of Sexuality: 2: The Use of Pleasure.* Penguin Books Limited.

Foucault M (2020) *The Foucault Reader: An Introduction to Foucault's Thought.* Penguin Books, Limited.

Foucault M (2020) *The History of Sexuality: 3: The Care of the Self.* Penguin Classics.

Foucault M (2020a) *The Foucault Reader: An Introduction to Foucault's Thought.* Penguin Books, Limited.

Foucault M (2020b) *The History of Sexuality: 1: The Will to Knowledge.* Penguin Books, Limited.

Foucault M (2020c) *Power: The Essential Works of Michel Foucault 1954-1984.* Penguin Books, Limited.

Foucault M (2020c) *The History of Sexuality: 3: The Care of the Self.* Penguin Classics.

Foucault M (2020d) *Power: The Essential Works of Michel Foucault 1954-1984.* Penguin Books, Limited.

Foucault M (2020e) *Society Must Be Defended: Lectures at the Collège de France, 1975-76.* Penguin Books, Limited.

Foucault M, Davidson AI and Burchell G (2010) *The Government of Self and Others: Lectures at the Collège de France 1982–1983.* Palgrave Macmillan UK.

Foucault, M. (2010) *The Government of Self and Others: Lectures at the Collège de France 1982–1983*, trans, Graham Burchell. New York: Palgrave Macmillan UK.

Fraser N (1985) Michel Foucault: A" young conservative"? *Ethics* 96(1): 165-184.

Fuchs M, Höpken W and Lexhagen M (2014) Big data analytics for knowledge generation in tourism destinations – A case from Sweden. *Journal of Destination Marketing & Management* 3(4): 198-209.

Galière S (2020) When food-delivery platform workers consent to algorithmic management: a Foucauldian perspective. *New Technology, Work and Employment* 35(3): 357-370.

Gangwar H, Date H and Raoot A (2014) Review on IT adoption: insights from recent technologies. *Journal of enterprise information management* 27(4): 488-502.

Ganti T (2014) Neoliberalism. *Annual Review of Anthropology* 43(1): 89-104.

Gardiner M (1996) Foucault, ethics and dialogue. *History of the Human Sciences* 9(3): 27-46.

Garland D (1997) `Governmentality' and the Problem of Crime::Foucault, Criminology, Sociology. *Theoretical Criminology* 1(2): 173-214.

Geiger RS (2017) Beyond opening up the black box: Investigating the role of algorithmic systems in Wikipedian organizational culture. *Big Data and Society* 4(2).

Geiger RS (2017) Beyond opening up the black box: Investigating the role of algorithmic systems in Wikipedian organizational culture. *Big Data and Society* 4(2).

Geis JR, Brady A, Wu CC, et al. (2019) Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. *Insights into Imaging* 10(1).

George G, Haas MR and Pentland A (2014) BIG DATA AND MANAGEMENT. *Academy of Management Journal* 57(2): 321-326.

George J and Wallio S (2017) Organizational justice and millennial turnover in public accounting. *Employee Relations* 39(1): 112-126.

Gibson C, Thomason B, Margolis J, et al. (2022) Dignity Inherent and Earned: The Experience of Dignity at Work. *Academy of Management Annals* 17(1): 218-267.

Gikopoulos J (2019) Alongside, not against: balancing man with machine in the HR function. *Strategic HR Review* 18(2): 56-61.

Gill R (2007) Postfeminist media culture: Elements of a sensibility. *European Journal of Cultural Studies* 10(2): 147-166.

Gill R (2008) *Discourse Analysis.* McGraw-Hill Companies.

Gollmitzer M (2023) Journalism ethics with Foucault: Casually employed journalists' constructions of professional integrity. *Journalism* 24(5): 1015-1033.

Graham C (2010) Accounting and the construction of the retired person. *Accounting, Organizations and Society* 35(1): 23-46.

Graham LJ (2011) The Product of Text and 'Other' Statements: Discourse analysis and the critical use of Foucault. *Educational Philosophy and Theory* 43(6): 663-674.

Graßmann C and Schermuly CC (2021) Coaching with artificial intelligence: concepts and capabilities. *Human Resource Development Review* 20(1): 106-126.

Gressgård R (2013) Asexuality: From pathology to identity and beyond. *Psychology and Sexuality* 4(2): 179-192.

Grey C (1996) Towards a critique of managerialism: The contribution of Simone Weil. *Journal of Management Studies* 33(5): 591-612.

Grønsund T and Aanestad M (2020) Augmenting the algorithm: Emerging human-in-the-loop work configurations. *The Journal of Strategic Information Systems* 29(2): 101614.

Guan C, Mou J and Jiang Z (2020) Artificial intelligence innovation in education: A twenty-year data-driven historical analysis. *International Journal of Innovation Studies* 4(4): 134-147.

Guba EG and Lincoln YS (1994) Competing paradigms in qualitative research. *Handbook of qualitative research.* Thousand Oaks, CA, US: Sage Publications, Inc, pp.105-117.

Guler B (2015) Innovations in information technology and the mortgage market. *Review of Economic Dynamics* 18(3): 456-483.

Günther WA, Rezazade Mehrizi MH, Huysman M, et al. (2017) Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems* 26(3): 191-209.

Hafermalz E (2021) Out of the panopticon and into exile: Visibility and control in distributed new culture organizations. *Organization Studies* 42(5): 697-717.

Hardy C and Thomas R (2014) Strategy, Discourse and Practice: The Intensification of Power. *Journal of Management Studies* 51(2): 320-348.

Hardy C and Thomas R (2015) Discourse in a Material World. *Journal of Management Studies* 52(5): 680.

Hargreaves T (2010) Putting Foucault to work on the environment: exploring pro-environmental behaviour change as a form of discipline. Reportno. Report Number|, Date. Place Published|: Institution|.

Hart C and Fuoli M (2020) Objectification strategies outperform subjectification strategies in military interventionist discourses. *Journal of Pragmatics* 162: 17-28.

Hartley J (2004) Essential Guide to Qualitative Methods in Organizational Research. London: SAGE Publications Ltd.

Hartley, J (2004) Case Study Research. In: Essential Guide to Qualitative Methods in Organizational Research. Jean Hartley Editor, London: SAGE Publications Ltd. pp. 323-333 Available at: <https://doi.org/10.4135/9781446280119> [Accessed 10 May 2024].

Hartmann K and Wenzelburger G (2021) Uncertainty, risk and the use of algorithms in policy decisions: a case study on criminal justice in the USA. *Policy Sciences* 54: 269-287.

Hazard GCJ (1990) The Future of Legal Ethics. *The Yale Law Journal* 100: 1239-1280.

Head E (2009) The ethics and implications of paying participants in qualitative research. *International Journal of Social Research Methodology* 12(4): 335-344.

Heiskala R (2001) Theorizing power: Weber, parsons, foucault and neostructuralism. *Social Science Information* 40(2): 241-264.

Heller KJ (1996) Power, subjectification and resistance in Foucault. *SubStance* 25(1): 78-110.

Hennink M, Hutter I and Bailey A (2020) *Qualitative Research Methods.* SAGE Publications.

Heßler PO, Pfeiffer J and Hafenbrädl S (2022) When Self-Humanization Leads to Algorithm Aversion. *Business & Information Systems Engineering* 64(3): 275-292.

Hildebrand-Nilshon M, Motzkau J and Papadopoulos D (2001). Reintegrating Sense into Subjectification. In: Morss, J.R., Stephenson, N., van Rappard, H. (eds) *Theoretical Issues in Psychology.* Springer, Boston, MA. https://doi.org/10.1007/978-1-4757-6817-6_25

Hill CJ and Lynn LE, Jr. (2004) Is Hierarchical Governance in Decline? Evidence from Empirical Research. *Journal of Public Administration Research and Theory* 15(2): 173-195.

Hill RK (2016) What an algorithm is. *Philosophy & Technology* 29: 35-59.

Hindess B (1996) *Discourses of Power: From Hobbes to Foucault.* Blackwell.

Hinkle GJ (1987) Foucault's power/knowledge and American sociological theorizing. *Human Studies*. 35-59.

Hodgson DE (2004) Project Work: The Legacy of Bureaucratic Control in the Post-Bureaucratic Organization. *Organization* 11(1): 81-100.

Hoffmann AL, Roberts ST, Wolf CT, et al. (2018) Beyond fairness, accountability, and transparency in the ethics of algorithms: Contributions and perspectives from LIS. *Proceedings of the Association for Information Science and Technology* 55(1): 694-696.

Hohendahl PU (1985) The dialectic of enlightenment revisited: Habermas' critique of the Frankfurt School. *New German Critique*.(35): 3-26.

Holford WD (2019) EMPHASIZING MÈTIS WITHIN THE DIGITAL ORGANIZATION. *Journal of Global Business and Technology* 15(1): 58-66.

Holford WD (2019) The future of human creative knowledge work within the digital economy. *Futures* 105: 143-154.

Howard JJ, Rabbitt LR and Sirotin YB (2020) Human-algorithm teaming in face recognition: How algorithm outcomes cognitively bias human decision-making. *Plos One* 15(8): e0237855.

Huber G (2022) Exercising power in autoethnographic vignettes to constitute critical knowledge. *Organization*. DOI: 10.1177/13505084221079006.

Huda M (2019) Empowering application strategy in the technology adoption. *Journal of Science and Technology Policy Management* 10(1): 172-192.

Introna LD (2016) Algorithms, Governance, and Governmentality: On Governing Academic Writing. *Science Technology and Human Values* 41(1): 17-49.

Isin E and Ruppert E (2020) The birth of sensory power: How a pandemic made it visible? *Big Data and Society* 7(2).

Jagatheesaperumal SK, Ahmad K, Al-Fuqaha A, et al. (2024) Advancing Education Through Extended Reality and Internet of Everything Enabled Metaverses: Applications, Challenges, and Open Issues. *IEEE Transactions on Learning Technologies* 17: 1120-1139.

Jago AS (2019) Algorithms and authenticity. *Academy of Management Discoveries* 5(1): 38-56.

Janssen M and Kuk G (2016) The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly* 33(3): 371-377.

Jarrahi MH (2018) Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons* 61(4): 577-586.

Jarrahi MH, Lutz C and Newlands G (2022) Artificial intelligence, human intelligence and hybrid intelligence based on mutual augmentation. *Big Data & Society* 9(2): 20539517221142824.

Jarrahi MH, Newlands G, Lee MK, et al. (2021) Algorithmic management in a work context. *Big Data & Society* 8(2): 20539517211020332.

Jauernig J, Uhl M and Walkowitz G (2022) People Prefer Moral Discretion to Algorithms: Algorithm Aversion Beyond Intransparency. *Philosophy & Technology* 35(1): 2.

Jia Q, Guo Y, Li R, et al. (2018) A conceptual artificial intelligence application framework in human resource management.

Jiang F, Jiang Y, Zhi H, et al. (2017) Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2(4): 230-243.

Jimenez-Anca JJ (2013) Beyond power: unbridging Foucault and Weber. *European Journal of Social Theory* 16(1): 36-50.

John C (2017) Will artificial intelligence usurp white collar jobs? *Human Resource Management International Digest* 25(3): 1-3.

Jöhnk J, Weißert M and Wyrtki K (2020) Ready or Not, AI Comes— An Interview Study of Organizational AI Readiness Factors. *Business & Information Systems Engineering* 63(1): 5-20.

Jones PE and Roelofsma PHMP (2000) The potential for social contextual and group biases in team decision-making: biases, conditions and psychological mechanisms. *Ergonomics* 43(8): 1129-1152.

Keller R (2011) The Sociology of Knowledge Approach to Discourse (SKAD). *Human Studies* 34(1): 43-65.

Keller R (2012) *Doing Discourse Research: An Introduction for Social Scientists.* SAGE Publications.

Kellogg KC, Valentine M and Christin A (2020) Algorithms at work: The new contested terrain of control. *The Academy of Management Annals* 14(1): 366-410.

Kelly M (1994) *Critique and Power: Recasting the Foucault/Habermas Debate.* MIT Press.

Khan, T. H., & MacEachen, E. (2021). Foucauldian Discourse Analysis: Moving Beyond a Social Constructionist Analytic. *International Journal of Qualitative Methods*, 20. https://doi.org/10.1177/16094069211018009

Kim B, Park J and Suh J (2020) Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. *Decision Support Systems* 134: 113302.

Kim K and Moon SI (2021) When Algorithmic Transparency Failed: Controversies Over Algorithm-Driven Content Curation in the South Korean Digital Environment. *American Behavioral Scientist* 65(6): 847-862.

Kivunja C and Kuyini AB (2017) Understanding and applying research paradigms in educational contexts. *International Journal of higher education* 6(5): 26-41.

Kleinberg J, Lakkaraju H, Leskovec J, et al. (2018) Human decisions and machine predictions. *The quarterly journal of economics* 133(1): 237-293.

Klumpp M (2018) Automation and artificial intelligence in business logistics systems: human reactions and collaboration requirements. *International Journal of Logistics-Research and Applications* 21(3): 224-242.

Knights D (2002) Writing Organizational Analysis into Foucault. *Organization* 9(4): 575-593.

Königs P (2020) Introduction to the Special Issue on the Ethics of State Mass Surveillance. *Moral Philosophy and Politics* 7(1): 1-8.

Kosmol T, Reimann F and Kaufmann L (2019) You'll never walk alone: Why we need a supply chain practice view on digital procurement. *Journal of Purchasing and Supply Management* 25(4): 100553.

Kotz DM (2002) Globalization and Neoliberalism. *Rethinking Marxism* 14(2): 64-79.

Kraemer F, van Overveld K and Peterson M (2010) Is there an ethics of algorithms? *Ethics and Information Technology* 13(3): 251-260.

Kroll JA, Huey J, Barocas S, et al. (2017) Accountable algorithms. *University of Pennsylvania Law Review* 165(3): 633-705.

Lange AC, Lenglet M and Seyfert R (2019) On studying algorithms ethnographically: Making sense of objects of ignorance. *Organization* 26(4): 598-617.

Langenbucher K (2020) Responsible AI-based credit scoring–a legal framework. *European Business Law Review* 31(4).

Langley A and Truax J (1994) A process study of new technology adoption in smaller manufacturing firms. *Journal of Management Studies* 31(5): 619-652.

Larner W (2000) Neo-liberalism: Policy, ideology, governmentality. *Studies in political economy* 63(1): 5-25.

Latour B (1987) *Science in Action: How to Follow Scientists and Engineers Through Society.* Harvard University Press.

Laundon M, Cathcart A and McDonald P (2019) Just benefits? Employee benefits and organisational justice. *Employee Relations: The International Journal* 41(4): 708-723.

Lauring J and Selmer J (2012) Knowledge sharing in diverse organisations. *Human Resource Management Journal* 22(1): 89-105.

Leclercq-Vandelannoitte A (2011) Organizations as discursive constructions: A Foucauldian approach. *Organization Studies* 32(9): 1247-1271.

Lee I and Shin YJ (2020) Machine learning for enterprises: Applications, algorithm selection, and challenges. *Business Horizons* 63(2): 157-170.

Lee N and Lings I (2008) *Doing Business Research: A Guide to Theory and Practice.* SAGE Publications.

Leicht-Deobald U, Busch T, Schank C, et al. (2019) The challenges of algorithm-based HR decision-making for personal integrity. *Journal of Business Ethics* 160(2): 377-392.

Leonardi PM (2021) COVID-19 and the new technologies of organizing: digital exhaust, digital footprints, and artificial intelligence in the wake of remote work. *Journal of Management Studies* 58(1): 249.

Leonardi PM and Treem JW (2020) Behavioral Visibility: A new paradigm for organization studies in the age of digitization, digitalization, and datafication. *Organization Studies* 41(12): 1601-1625.

Lepri B, Oliver N, Letouzé E, et al. (2018) Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy and Technology* 31(4): 611-627.

Lepri B, Staiano J, Sangokoya D, et al. (2017) The Tyranny of Data? The Bright and Dark Sides of Data-Driven Decision-Making for Social Good. *Studies in Big Data*. pp.3-24.

Leslie D (2019) Understanding artificial intelligence ethics and safety.

Lewis MW (2000) Exploring Paradox: Toward a More Comprehensive Guide. *Academy of Management Review* 25(4): 760-776.

Li F, Ruijs N and Lu Y (2022) Ethics & AI: A systematic review on ethical concerns and related strategies for designing with AI in healthcare. *AI* 4(1): 28-53.

Lilja M and Vinthagen S (2014) Sovereign power, disciplinary power and biopower: resisting what power with what resistance? *Journal of political power* 7(1): 107-126.

Lincoln YS and Guba EG (1985) *Naturalistic Inquiry.* SAGE Publications.

Lindebaum D, Vesa M and Den Hond F (2020) Insights from “the machine stops” to better understand rational assumptions in algorithmic decision making and its implications for organizations. *Academy of Management Review* 45(1): 247-263.

Lofland J, Snow D, Anderson L, et al. (2022) *Analyzing Social Settings: A Guide to Qualitative Observation and Analysis, Fourth Edition.* Waveland Press.

Logg JM, Minson JA and Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151: 90-103.

Love NS (1989) Foucault & Habermas on discourse & democracy. *Polity* 22(2): 269-293.

Lucas K (2015) Workplace Dignity: Communicating Inherent, Earned, and Remediated Dignity. *Journal of Management Studies* 52(5): 621-646.

Lyon D (1993) An electronic panopticon? A sociological critique of surveillance theory. *The Sociological Review* 41(4): 653-678.

MacDonald MN, Badger R and Dasli M (2006) Authenticity, culture and language learning. *Language and Intercultural Communication* 6(3-4): 250-261.

Magalhães JC (2018) Do algorithms shape character? Considering algorithmic ethical subjectivation. *Social Media+ Society* 4(2): 2056305118768301.

Makarius EE, Mukherjee D, Fox JD, et al. (2020) Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research* 120: 262-273.

Manfreda A, Ljubi K and Groznik A (2019) Autonomous vehicles in the smart city era: An empirical study of adoption factors important for millennials. *International Journal of Information Management*. DOI: https://doi.org/10.1016/j.ijinfomgt.2019.102050. 102050.

Mansoury M, Abdollahpouri H, Pechenizkiy M, et al. (2020) Feedback loop and bias amplification in recommender systems. *Proceedings of the 29th ACM international conference on information & knowledge management.* 2145-2148.

Martin K (2022) Value-laden Biases in Data Analytics. *Ethics of Data and Analytics*. Auerbach Publications, pp.1-5.

Marx GT (1998) Ethics for the New Surveillance. *The Information Society* 14(3): 171-185.

Mason J (2017) *Qualitative Researching.* SAGE Publications.

Matthewman S (2013) Michel Foucault, technology, and actor-network theory. *Techné: Research in Philosophy and Technology* 17(2): 274-292.

McHoul A, McHoul A, Wendy Grace both of Murdoch University MA, et al. (2015) *A Foucault Primer: Discourse, Power And The Subject.* Taylor & Francis.

McHoul AW and Grace W (1998) *A Foucault Primer: Discourse, Power, and the Subject.* Otago University Press.

McKay C (2020) Predicting risk in criminal procedure: actuarial tools, algorithms, AI and judicial decision-making. *Current Issues in Criminal Justice* 32(1): 22-39.

McKinlay A and Starkey K (1998) *Foucault, Management and Organization Theory: From Panopticon to Technologies of Self.* SAGE Publications.

McMullan T (2015) *What does the panopticon mean in the age of digital surveillance?* Available at: https://www.theguardian.com/technology/2015/jul/23/panopticon-digital-surveillance-jeremy-bentham [Accessed: 24 February 2025].

McMurray R, Pullen A and Rhodes C (2011) Ethical subjectivity and politics in organizations: A case of health care tendering. *Organization* 18(4): 541-561.

Meijer AJ (2008) E-mail in government: Not post-bureaucratic but late-bureaucratic organizations. *Government Information Quarterly* 25(3): 429-447.

Mennicken A and Miller P (2014) Michel Foucault and the Administering of Lives. In: Adler P, du Gay P, Morgan G, et al. (eds) *The Oxford Handbook of Sociology, Social Theory, and Organization Studies: Contemporary Currents*. Oxford University Press, pp.0.

Metcalf L, Askay DA and Rosenberg LB (2019) Keeping Humans in the Loop: Pooling Knowledge through Artificial Swarm Intelligence to Improve Business Decision Making. *California Management Review* 61(4): 84-109.

Milano S, Taddeo M and Floridi L (2020) Recommender systems and their ethical challenges. *Ai & Society* 35(4): 957-967.

Miles MB, Huberman AM and Saldana J (2018) *Qualitative Data Analysis: A Methods Sourcebook.* SAGE Publications.

Mills S (2003) *Michel Foucault.* Routledge.

Mingers J and Willcocks LP (2004) *Social Theory and Philosophy for Information Systems.* Wiley.

Minocher X and Randall C (2020) Predictable policing: New technology, old bias, and future resistance in big data surveillance. *Convergence-the International Journal of Research into New Media Technologies* 26(5-6): 1108-1124.

Miron-Spektor E, Ingram A, Keller J, et al. (2018) Microfoundations of Organizational Paradox: The Problem Is How We Think about the Problem. *Academy of Management Journal* 61(1): 26-45.

Mittelstadt BD, Allo P, Taddeo M, et al. (2016) The ethics of algorithms: Mapping the debate. *Big Data & Society* 3(2): 1-21.

Mogaji E and Nguyen NP (2022) Managers' understanding of artificial intelligence in relation to marketing financial services: insights from a cross-country study. *International Journal of Bank Marketing* 40(6): 1272-1298.

Moisander J, Groß C and Eräranta K (2018) Mechanisms of biopower and neoliberal governmentality in precarious work: Mobilizing the dependent self-employed as independent business owners. *Human Relations* 71(3): 375-398.

Moisander J, Groß C and Eräranta K (2018) Mechanisms of biopower and neoliberal governmentality in precarious work: Mobilizing the dependent self-employed as independent business owners. *Human Relations* 71(3): 375-398.

Moran RE and Shaikh SJ (2022) Robots in the news and newsrooms: Unpacking meta-journalistic discourse on the use of artificial intelligence in journalism. *Digital Journalism* 10(10): 1756-1774.

Moulaison HL, Dykas F and Budd JM (2014) Foucault, the author, and intellectual debt: Capturing the author-function through attributes, relationships, and events in knowledge organization systems. *Knowledge Organization* 41(1): 30-43.

Mujtaba DF and Mahapatra NR (2019) Ethical Considerations in AI-Based Recruitment. IEEE, 1-7.

Munro I (2014) Organizational Ethics and Foucault's 'Art of Living': Lessons from Social Movement Organizations. *Organization Studies* 35(8): 1127-1148.

Munro I (2017) Whistle-blowing and the politics of truth: Mobilizing 'truth games' in the WikiLeaks case. *Human Relations* 70(5): 519-543.

Munro I (2018) An interview with Snowden's lawyer: Robert Tibbo on whistleblowing, mass surveillance and human rights activism. *Organization* 25(1): 106-122.

Murphy MH (2017) Algorithmic surveillance: the collection conundrum. *International Review of Law, Computers & Technology* 31(2): 225-242.

Murray J and Flyverbom M (2020) Datafied corporate political activity: Updating corporate advocacy for a digital era. *Organization* 28(4): 621-640.

Newbutt N, Schmidt MM, Riva G, et al. (2020) The possibility and importance of immersive technologies during COVID-19 for autistic people. *Journal of Enabling Technologies* 14(3): 187-199.

Newlands G (2021) Algorithmic surveillance in the gig economy: The organization of work through Lefebvrian conceived space. *Organization Studies* 42(5): 719-737.

Neyland D (2015) On organizing algorithms. *Theory, Culture & Society* 32(1): 119-132.

Neyland D (2016) Bearing Account-able Witness to the Ethical Algorithmic System. *Science, Technology, & Human Values* 41(1): 50-76.

Neyland D and Möllers N (2016) Algorithmic IF … THEN rules and the conditions and consequences of power. *Information, Communication & Society* 20(1): 45-62.

Nigam A (1996) Marxism and Power. *Social Scientist* 24(4/6): 3-22.

Nilashi M and Abumalloh RA (2024) i-TAM: A model for immersive technology acceptance. *Education and Information Technologies*. DOI: 10.1007/s10639-024-13080-5.

Nissen M (2003) Objective subjectification: The antimethod of social work. *Mind, Culture, and Activity* 10(4): 332-349.

Nola R (1998). Knowledge, discourse, power and genealogy in Foucault. *Critical Review of International Social and Political Philosophy,* 1(2), 109–154. https://doi.org/10.1080/13698239808403240

Nussbaum MC (1995) Objectification. *Philosophy & Public Affairs* 24(4): 249-291.

O'Neil C (2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Penguin Books Limited.

O'Farrell C (2005) *Michel Foucault.* SAGE Publications.

Oakes LS, Townley B and Cooper DJ (1998) Business planning as pedagogy: Language and control in a changing institutional field. *Administrative Science Quarterly*. 257-292.

Okamura K and Yamada S (2020) Adaptive trust calibration for human-AI collaboration. *Plos One* 15(2).

Oliveira T and Martins MF (2011) Literature review of information technology adoption models at firm level. *Electronic journal of information systems evaluation* 14(1): 110-121.

Olssen M (2003) Structuralism, post-structuralism, neo-liberalism: assessing Foucault's legacy. *Journal of Education Policy* 18(2): 189-202.

Orea-Giner A, Muñoz-Mazón A, Villacé-Molinero T, et al. (2022) Cultural tourist and user experience with artificial intelligence: a holistic perspective from the Industry 5.0 approach. *Journal of Tourism Futures*. DOI: 10.1108/JTF-04-2022-0115.

Oswald FL, Behrend TS, Putka DJ, et al. (2020) Big Data in Industrial-Organizational Psychology and Human Resource Management: Forward Progress for

Organizational Research and Practice. *Annual Review of Organizational Psychology and Organizational Behavior* 7(1): 505-533.

Oswald M, Grace J, Urwin S, et al. (2018) Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental'proportionality. *Information & Communications Technology Law* 27(2): 223-250.

Ouchchy L, Coin A and Dubljević V (2020) AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. *AI & SOCIETY: Journal of Knowledge, Culture and Communication* 35(4): 927.

Overall J (2019) The Ethics of Mass Surveillance: An Anarchist, Objectivist, and Critical Theorist Perspective. *Journal of Information Ethics* 28(2): 34-50.

Ozanne LK, Phipps M, Weaver T, et al. (2016) Managing the Tensions at the Intersection of the Triple Bottom Line: A Paradox Theory Approach to Sustainability Management. *Journal of Public Policy & Marketing* 35(2): 249-261.

Pachankis Y (2022) Mass Surveillance, Behavioural Control, and Psychological Coercion the Moral Ethical Risks in Commercial Devices. *Computer Science & Information Technology* 12(13): 151-168.

Pachidi S, Berends H, Faraj S, et al. (2021) Make Way for the Algorithms: Symbolic Actions and Change in a Regime of Knowing. *Organization Science* 32(1): 18.

Peeters R (2020) The agency of algorithms: Understanding human-algorithm interaction in administrative decision-making. *Information Polity* 25(4): 507-522.

Peña-Acuña B and Rubio-Alcalá FD (2024) Ethical approach to the use of immersive technologies. Advance about digitalisation of multilingual programs in the EHEA. *Frontiers in Virtual Reality* 5.

Perrow C (1979) *Complex Organizations: A Critical Essay.* Scott, Foresman.

Pfeffer J (1992) Understanding power in organizations. *California Management Review* 34(2): 29-50.

Phillips N and Hardy C (2002) *Discourse Analysis: Investigating Processes of Social Construction.* SAGE Publications.

Phillips RA and Margolis JD (1999) Toward an Ethics of Organizations. *Business Ethics Quarterly* 9(4): 619-638.

Pieters WR (2018) Assessing organisational justice as a predictor of job satisfaction and employee engagement in Windhoek. *SA Journal of Human Resource Management* 16(1): 1-11.

Pignot E (2021) Who is pulling the strings in the platform economy? Accounting for the dark and unexpected sides of algorithmic control. *Organization*. DOI: 10.1177/1350508420974523.

Pillai R and Sivathanu B (2020) Adoption of artificial intelligence (AI) for talent acquisition in IT/ITeS organizations. *Benchmarking* 27(9): 2599-2629.

Pimentel D, Brown N, Vecchio F, et al. (1992) Ethical issues concerning potential global climate change on food production. *Journal of Agricultural and Environmental Ethics* 5(2): 113-146.

Prem, E. (2023) From ethical AI frameworks to tools: a review of approaches. *AI and Ethics* 3, 699–716. https://doi.org/10.1007/s43681-023-00258-9.

Prikshat V, Malik A and Budhwar P (2023) AI-augmented HRM: Antecedents, assimilation and multilevel consequences. *Human Resource Management Review* 33(1).

Punch KF (2013) *Introduction to Social Research: Quantitative and Qualitative Approaches.* SAGE Publications.

Raffnsøe S, Gudmand-Høyer M and Thaning MS (2016) Foucault's dispositive: The perspicacity of dispositive analytics in organizational research. *Organization* 23(2): 272-298.

Raffnsøe S, Mennicken A and Miller P (2019) The Foucault Effect in Organization Studies. *Organization Studies* 40(2): 155-182.

Ratten V (2012) Entrepreneurial and ethical adoption behaviour of cloud computing. *The Journal of High Technology Management Research* 23(2): 155-164.

Redden J (2018) Democratic governance in an age of datafication: Lessons from mapping government discourses and practices. *Big Data and Society* 5(2).

Reinares-Lara E, Olarte-Pascual C and Pelegrín-Borondo J (2018) Do you want to be a cyborg? The moderating effect of ethics on neural implant acceptance. *Computers in Human Behavior* 85: 43-53.

Ritchie J, Lewis J, Lewis PSPJ, et al. (2013) *Qualitative Research Practice: A Guide for Social Science Students and Researchers.* SAGE Publications.

Roberts H, Cowls J, Morley J, et al. (2021) The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation. *Ai & Society* 36(1): 59-77.

Roberts K and Herrington V (2013) Organisational and procedural justice: a review of the literature and its implications for policing. *Journal of Policing, Intelligence and Counter Terrorism* 8(2): 115-130.

Roberts SL (2019) Big data, algorithmic governmentality and the regulation of pandemic risk. *European Journal of Risk Regulation* 10(1): 94-115.

Robillard JM, Cleland I, Hoey J, et al. (2018) Ethical adoption: A new imperative in the development of technology for dementia. *Alzheimer's & Dementia* 14(9): 1104-1113.

Robinson SC (2020) Trust, transparency, and openness: How inclusion of cultural values shapes Nordic national public policy strategies for artificial intelligence (AI). *Technology in Society* 63: 101421.

Robson C and McCartan K (2016) *Real World Research.* Wiley.

Rogers EM (2003) *Diffusion of Innovations, 5th Edition.* Free Press.

Roßmann B, Canzaniello A, von der Gracht H, et al. (2018) The future and social impact of Big Data Analytics in Supply Chain Management: Results from a Delphi study. *Technological Forecasting and Social Change* 130: 135-149.

Rouse J (2005) Power/Knowledge. In: Gutting G (ed) *The Cambridge Companion to Foucault*. 2 ed. Cambridge: Cambridge University Press, pp.95-122.

Ruppert ES (2008) 'I is; therefore i am': The census as practice of double identification. *Sociological Research Online* 13(4).

Russell S, Hauert S, Altman R, et al. (2015) Robotics: Ethics of artificial intelligence. *Nature* 521(7553): 415-418.

Russo F, Schliesser E and Wagemans J (2024) Connecting ethics and epistemology of AI. *Ai & Society* 39(4): 1585-1603.

Safdar NM, Banja JD and Meltzer CC (2020) Ethical considerations in artificial intelligence. *European Journal of Radiology* 122.

Sagiroglu S and Sinanc D (2013) Big data: A review. *2013 international conference on collaboration technologies and systems (CTS).* IEEE, 42-47.

Salter MB (2019) Security actor-network theory: Revitalizing securitization theory with Bruno Latour. *Polity* 51(2): 349-364.

Sánchez Laws AL and Utne T (2019) Ethics Guidelines for Immersive Journalism. *Frontiers in Robotics and Ai* 6.

Saunders B, Kitzinger J and Kitzinger C (2015a) Anonymising interview data: challenges and compromise in practice. *Qualitative Research* 15(5): 616-632.

Saunders M, Lewis P and Thornhill A (2015b) *Research Methods for Business Students.* Pearson Education.

Sayer A (2007) Dignity at Work: Broadening the Agenda. *Organization* 14(4): 565-581.

Schniter E, Shields TW and Sznycer D (2020) Trust in humans and robots: Economically similar but emotionally different. *Journal of Economic Psychology* 78: 102253.

Schoonhoven CB (1981) Problems with Contingency Theory: Testing Assumptions Hidden within the Language of Contingency "Theory". *Administrative Science Quarterly* 26(3): 349-377.

Schraub D (2016) Racism as Subjectification. *Berkeley Journal of African-American Law & Policy Sciences* 17: 3-46.

Schwandt TA (1994) Constructivist, interpretivist approaches to human inquiry. *Handbook of qualitative research.* Thousand Oaks, CA, US: Sage Publications, Inc, pp.118-137.

Schwerzmann K (2021) Abolish! Against the use of risk assessment algorithms at sentencing in the US Criminal Justice System. *Philosophy & Technology* 34(4): 1883-1904.

Seeber I, Bittner E, Briggs RO, et al. (2020) Machines as teammates: A research agenda on AI in team collaboration. *Information & Management* 57(2): 103174.

Seeck H (2011) Exploring the Foucauldian interpretation of power and subject in organizations. *Journal of Management and Organization* 17(6).

Seeck H and Kantola A (2009) Organizational control: Restrictive or productive? *Journal of Management and Organization* 15(2).

Sestino A and De Mauro A (2022) Leveraging Artificial Intelligence in Business: Implications, Applications and Methods. *Technology Analysis & Strategic Management* 34(1): 16-29.

Shah H (2018) Algorithmic accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376(2128).

Shaw D and Scully J (2023) The foundations of influencing policy and practice: How risk science discourse shaped government action during COVID-19. *Risk Analysis* 0(0): 1-17.

Sheehan B, Jin HS and Gottlieb U (2020) Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research* 115: 14-24.

Sherwani F, Asad MM and Ibrahim BSKK (2020) Collaborative Robots and Industrial Revolution 4.0 (IR 4.0). *2020 International Conference on Emerging Trends in Smart Technologies, ICETST 2020.*

Siau K and Wang W (2020) Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management (JDM)* 31(2): 74-87.

Silverman D (2009) *Doing Qualitative Research.* SAGE Publications.

Simmler M, Brunner S, Canova G, et al. (2023) Smart criminal justice: exploring the use of algorithms in the Swiss criminal justice system. *Artificial Intelligence and Law* 31(2): 213-237.

Simon J (1994) Between power and knowledge: Habermas, foucault, and the future of legal studies: Comment. *Law & Society Review* 28(4): 947-961.

Simon O, Neuhofer B and Egger R (2020) Human-robot interaction: Conceptualising trust in frontline teams through LEGO® Serious Play®. *Tourism Management Perspectives* 35: 100692.

Skeem J and Lowenkamp C (2020) Using algorithms to address trade-offs inherent in predicting recidivism. *Behavioral Sciences & the Law* 38(3): 259-278.

Skinner D (2013) Foucault, subjectivity and ethics: Towards a self-forming subject. *Organization* 20(6): 904-923.

Smart C (1992) The Woman of Legal Discourse*. Social & Legal Studies,* 1(1), 29-44. https://doi.org/10.1177/096466399200100103

Smith WK and Lewis MW (2011) Toward a Theory of Paradox: A Dynamic equilibrium Model of Organizing. *Academy of Management Review* 36(2): 381-403.

Soliman MM, Ahmed E, Darwish A, et al. (2024) Artificial intelligence powered Metaverse: analysis, challenges and future perspectives. *Artificial Intelligence Review* 57(2).

Southgate E, Smith SP, Cividino C, et al. (2019) Embedding immersive virtual reality in classrooms: Ethical, organisational and educational lessons in bridging research and practice. *International Journal of Child-Computer Interaction* 19: 19-29.

Spencer AJ, Buhalis D and Moital M (2012) A hierarchical model of technology adoption for small owner-managed travel firms: An organizational decision-making and leadership perspective. *Tourism Management* 33(5): 1195-1208.

Stamate AN, Sauvé G and Denis PL (2021) The rise of the machines and how they impact workers' psychological health: An empirical study. *Human Behavior and Emerging Technologies* 3(5): 942-955.

Staunæs D (2003) Where have all the subjects gone? Bringing together the concepts of intersectionality and subjectification. *NORA: Nordic journal of women's studies* 11(2): 101-110.

Strauss A and Corbin JM (1990) *Basics of Qualitative Research: Grounded Theory Procedures and Techniques.* SAGE Publications.

Suh A and Prophet J (2018) The state of immersive technology research: A literature analysis. *Computers in Human Behavior* 86: 77-90.

Sun TQ and Medaglia R (2019) Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly* 36(2): 368-383.

Sun W, Nasraoui O and Shafto P (2020) Evolution and impact of bias in human and machine learning algorithm interaction. *Plos One* 15(8): e0235502.

Sung EC, Bae S, Han DID, et al. (2021) Consumer engagement via interactive artificial intelligence and mixed reality. *International Journal of Information Management* 60.

Suseno Y, Chang C, Hudik M, et al. (2021) Beliefs, anxiety and change readiness for artificial intelligence adoption among human resource managers: the moderating role of high-performance work systems. *The International Journal of Human Resource Management* 33(6): 1209-1236.

Susse T, Kobert M and Kries C (2021) Antecedents of Constructive Human-AI Collaboration: An Exploration of Human Actors' Key Competencies. In: *SMART AND SUSTAINABLE COLLABORATIVE NETWORKS 4.0 (PRO-VE 2021)*, pp.113-124.

Sutton SG, Arnold V and Holt M (2018) How Much Automation Is Too Much? Keeping the Human Relevant in Knowledge Work. *Journal of Emerging Technologies in Accounting* 15(2): 15-25.

Syed R (2020) Cybersecurity vulnerability management: A conceptual ontology and cyber intelligence alert system. *Information & Management* 57(6): 103334.

Taddeo M and Floridi L (2018) How AI can be a force for good. *Science (New York, N.Y.)* 361(6404): 751-752.

Tambe P, Cappelli P and Yakubovich V (2019) Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review* 61(4): 15-42.

Tarafdar M, Page X and Marabelli M (2023) Algorithms as co-workers: Human algorithm role interactions in algorithmic work. *Information Systems Journal* 33(2): 232-267.

Tom Dieck MC and Han DID (2022) The role of immersive technology in Customer Experience Management. *Journal of Marketing Theory and Practice* 30(1): 108-119.

Tornatzky LG, Fleischer M and Chakrabarti AK (1990) *The Processes of Technological Innovation.* Lexington Books.

Törnroos M, Elovainio M, Hintsa T, et al. (2019) Personality traits and perceptions of organisational justice. *International Journal of Psychology* 54(3): 414-422.

Townley B (1993) Foucault, Power/Knowledge, and Its Relevance for Human Resource Management. *The Academy of Management Review* 18(3): 518-545.

Townley B (1995) Know thyself': Self-awareness, self-formation and managing. *Organization* 2(2): 271-289.

Trendall S (2023) Government guidance bans civil servants from using ChatGPT to write policy papers. Available at: https://www.publictechnology.net/2023/07/03/science-technology-and-research/government-guidance-bans-civil-servants-from-using-chatgpt-to-write-policy-papers (Accessed: 3 August 2024).

Tsai C-Y, Marshall JD, Choudhury A, et al. (2022) Human-robot collaboration: A multilevel and integrated leadership framework. *The Leadership Quarterly* 33(1): 101594.

Tsamados A, Aggarwal N, Cowls J, et al. (2022) The ethics of algorithms: key problems and solutions. *Ai & Society* 37(1): 215-230.

Tsamados A, Aggarwal N, Cowls J, et al. (2022) The ethics of algorithms: key problems and solutions. *Ai & Society* 37(1): 215-230.

Tucker J (1993) Everyday forms of employee resistance. *Sociological Forum* 8(1): 25-45.

Tursunbayeva A, Pagliari C, Di Lauro S, et al. (2022) The ethics of people analytics: risks, opportunities and recommendations. *Personnel Review* 51(3): 900-921.

Tweedale JW (2013) Using multi-agent systems to pursue autonomy with automated components. *Procedia Computer Science.* 1369-1378.

Upchurch M (2018) Robots and AI at work: the prospects for singularity. *New Technology, Work and Employment* 33(3): 205-218.

Uphoff N (1989) Distinguishing power, authority & legitimacy: Taking Max Weber at his word by using resources-exchange analysis. *Polity* 22(2): 295-322.

Ussher JM and Perz J (2020) “I feel fat and ugly and hate myself”: Self-objectification through negative constructions of premenstrual embodiment. *Feminism and Psychology* 30(2): 185-205.

Van Dijk R and Van Dick R (2009) Navigating Organizational Change: Change Leaders, Employee Resistance and Work-based Identities. *Journal of Change Management* 9(2): 143-163.

Van Dijk TA (1985) *Handbook of Discourse Analysis: Discourse analysis in society.* Academic Press.

Van Esch P and Black JS (2019) Factors that influence new generation candidates to engage with and complete digital, AI-enabled recruiting. *Business Horizons* 62(6): 729-739.

Vassilopoulou J, Kyriakidou O, Özbilgin MF, et al. (2022) Scientism as illusio in HR algorithms: Towards a framework for algorithmic hygiene for bias proofing. *Human Resource Management Journal*. DOI: 10.1111/1748-8583.12430.

Velkova J and Kaun A (2021) Algorithmic resistance: Media practices and the politics of repair. *Information, Communication & Society* 24(4): 523-540.

Waardenburg L, Huysman M and Sergeeva AV (2022) In the Land of the Blind, the One-Eyed Man Is King: Knowledge Brokerage in the Age of Learning Algorithms. *Organization Science* 33(1): 59-82.

Walker M, Fleming P and Berti M (2021) 'You can't pick up a phone and talk to someone': How algorithms function as biopower in the gig economy. *Organization* 28(1): 26-43.

Weber M (1978) *Economy and Society: An Outline of Interpretive Sociology.* University of California Press.

Wei Z and Yuan M (2023) Research on the Current Situation and Future Development Trend of Immersive Virtual Reality in the Field of Education. *Sustainability*, 15.

Weiskopf R and Hansen HK (2023) Algorithmic governmentality and the space of ethics: Examples from 'People Analytics'. *Human Relations* 76(3): 483-506.

Weiskopf R and Willmott H (2013) Ethics as Critical Practice: The "Pentagon Papers", Deciding Responsibly, Truth-telling, and the Unsettling of Organizational Morality. *Organization Studies* 34(4): 469-493.

Wheeler B (2020) Reliabilism and the Testimony of Robots. *Techné: Research in Philosophy and Technology* 24(3): 332-356.

White R (2014) Foucault on the Care of the Self as an Ethical Project and a Spiritual Goal. *Human Studies* 37(4): 489-504.

Wiggershaus R (1994) *The Frankfurt School: Its History, Theories, and Political Significance.* MIT Press.

Wolf CT and Blomberg JL (2019) Evaluating the promise of human-algorithm collaborations in everyday work practices. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW).

Wood AW (1991) Unsociable sociability: The anthropological basis of Kantian ethics. *Philosophical topics* 19(1): 325-351.

Xi M, Perera M, Matthews B, et al. (2024) Towards Immersive AI. *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct).* 260-264.

Yin RK (2018) *Case study research and applications : design and methods.* SAGE.

Yu H (2023) Reflection on whether Chat GPT should be banned by academia from the perspective of education and teaching. *Frontiers in Psychology* 14.

Yu H, Shen Z, Miao C, et al. (2018) Building Ethics into Artificial Intelligence. In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18), 2018*, pp.5527-5533.

Završnik A (2021) Algorithmic justice: Algorithms and big data in criminal justice settings. *European Journal of criminology* 18(5): 623-642.

Zheng NN, Liu ZY, Ren PJ, et al. (2017) Hybrid-augmented intelligence: collaboration and cognition. *Frontiers of Information Technology & Electronic Engineering* 18(2): 153-179.

Zhou Y, Zhang X and Ding F (2022) Partially-coupled nonlinear parameter optimization algorithm for a class of multivariate hybrid models. *Applied Mathematics and Computation* 414: 126663.

Zhu K and Weyant JP (2003) Strategic decisions of new technology adoption under asymmetric information: a game-theoretic model. *Decision sciences* 34(4): 643-675.

Zuboff S (2019) *The Age of Surveillance Capitalism: The fight for a human future at the new frontier of power.* New York: PublicAffairs.

APPENDIX 1: Participant Information Sheet

Aston Business School
BIRMINGHAM UK

Aston Business School
Aston University
Birmingham B4 7ET
United Kingdom

+44 (0)121 204 3000
aston.ac.uk/abs

Participant Information Sheet

Study title

An exploration of strategic decision-making around the adoption and implementation of Algorithmic platforms into organisations/working environments from the perspectives of key organisational stakeholders.

What is the purpose of this study?

Our study seeks to understand how organisational contextual factors influence the discussion and decision-making around the introduction and implementation of algorithmic technologies. We have been informed that [redacted] has been utilizing algorithmic prediction tools for [redacted] (e.g., [redacted] *tools)*. Thus, we are particularly interested in investigating the motivations for adopting these algorithmic systems, the key decision-makers in this process, and the main factors in the design and implementation of such platforms at [redacted] work practices. We are also interested in investigating how effectively this tool has been introduced in your organisation and the impacts this has had on work performance.

What will happen if I take part?

We wish to speak to different employees across the organisation about their perceptions of, and work experiences with, algorithmic prediction systems.

If you are interested in taking part, we will contact you to set-up a date for our interview. Interviews will usually take no longer than 30-40 minutes and we will look to timetable them in a way that will be the least disruptive for yourself and your work. We can hold these interviews either face-to-face in a private location at your workplace or via an online platform of your choice (e.g., Teams or Zoom). With your agreement, we would like to record the interview, so we can more accurately recall its content. However, if you would like to proceed without our recording the interview, that is also OK.

Once the interview is completed, we will send you a £10 Amazon e-voucher as a token of our gratitude for your support with this research. You will receive this via your organisation email.

What are the potential benefits of taking part?

Our research seeks to evaluate the decision-making processes that underpin the introduction of AI / algorithmic assessment systems [redacted] workplace.

TRIPLE ACCREDITED AACSB ACCREDITED EQUIS ACCREDITED ASSOCIATION OF MBAS ACCREDITED

Aston Business School
BIRMINGHAM UK

Aston Business School
Aston University
Birmingham B4 7ET
United Kingdom

+44 (0)121 204 3000
aston.ac.uk/abs

As such, we hope our research will help to inform, and improve, future decision-making processes regarding the introduction of algorithms, AI technologies and other intelligent predictive tools. By participating in this research, therefore, you are providing an important voice in this process, helping [redacted] to further improve decision-making around the introduction of algorithmic technologies into your work practices.

Why have I been invited to take part?

You have been invited to take part for one of two reasons. First, your work directly or indirectly interacts with the algorithmic tools of [redacted]. Second, you have been involved in decisions to introduce algorithms into Risk Assessment practices. Consequently, we feel you have an important insight into the adoption, introduction, implementation, and/or management of algorithmic technologies in [redacted] organisational working practices.

Data Protection and Confidentiality

Aston University take its obligations under data ad privacy law seriously, and complies with its Data Protection Policies and Legislations, its Record Management Policy and Procedures and the University's Information Security Policy. Furthermore, all data produced by this research will be retained in accordance with Aston University's policy on Academic Integrity. The University will also ensure that the data is collected as part of the research study will be kept confidential and will only be used for research and academic purposes.

For the purposes of the research study, the data that will be obtained will be pseudonymised. This means that the personal data you provide will be replaced with a pseudonym i.e., a value or a code which does not allow your data to be directly identified.

The University, in accordance with data and privacy law, will ensure that the personal data provided is stored separately to the pseudonymised data. In accordance with its Information Security Policy, we will ensure that appropriate measures are adopted to ensure data is stored securely. Data will be stored in Aston University's Box storage system.

The University will comply with its obligations under data and privacy laws and ensure that relevant consents are obtained and that it stores personal data in accordance with its policies and procedures.

What will happen to the results of this research project?

The results of this research will be used for the primary purpose of completing a PhD thesis. The results will also be used to write academic paper(s), which will be written and published

Aston Business School
BIRMINGHAM UK

Aston Business School
Aston University
Birmingham B4 7ET
United Kingdom

+44 (0)121 204 3000
aston.ac.uk/abs

in the public domain for research and educational purposes. Participating organisations or individuals will not be identifiable and will remain anonymous within all these publications.

Who is organising and funding this research?

Ali Gordjahanbeiglou is conducting this research as PhD candidate at Aston University Business School, under the supervision of Dr Jonathan Crawshaw, Prof Nicholas Theodorakopoulos and Dr Judy Scully (external supervisor). This project is self-funded.

Who has reviewed the study?

The study has been approved by University Research Ethics Committee of Aston University as well as [redacted]-Review Board (**Ref:** 2022-026). Please refer to the appendix for [redacted] final approval letter.

Contact for Further Information

Project Lead: Ali Gordjahanbeiglou (PhD candidate) agord19@aston.ac.uk

Lead supervisor; Dr Jonathan Crawshaw (j.r.crawshaw2@aston.ac.uk).

Prof Nicholas Theodorakopoulos

Dr Judy Scully

If you have any concern about the way in which the study has been conducted, you can contact Secretary of the Aston Business School Research Ethics Committee on: jonesp5@aston.ac.uk or abs_aarm@aston.ac.uk .

Thank you for taking time to read this information sheet **Date: March 2023**

Aston Business School
BIRMINGHAM UK

Aston Business School
Aston University
Birmingham B4 7ET
United Kingdom

+44 (0)121 204 3000
aston.ac.uk/abs

Appendix or Links: *Aston University takes its obligations under data and privacy law seriously and complies with the General Data Protection Regulation ("GDPR") and the Data Protection Act 2018 ("DPA"). Aston University is the sponsor for this study based in the United Kingdom. We will be using information from you in order to undertake this study. Aston University will process your personal data in order to register you as a participant and to manage your participation in the study. It will process your personal data on the grounds that it is necessary for the performance of a task carried out in the public interest (GDPR Article 6(1)(e).*

Aston University may process special categories of data about you which includes details about your health. Aston University will process this data on the grounds that it is necessary for statistical or research purposes (GDPR Article 9(2)(j)). Aston University will keep identifiable information about you for 6 years after the study has finished.

Your rights to access, change or move your information are limited, as we need to manage your information in specific ways in order for the research to be reliable and accurate. If you withdraw from the study, you have the right to withdraw your data. To safeguard your rights, we will use the minimum personally identifiable information possible.

You can find out more about how we use your information at www.aston.ac.uk/dataprotection or by contacting our Data Protection Officer at dp_officer@aston.ac.uk .

If you wish to raise a complaint on how we have handled your personal data, you can contact our Data Protection Officer who will investigate the matter. If you are not satisfied with our response or believe we are processing your personal data in a way that is not lawful you can complain to the Information Commissioner's Office (ICO).

TRIPLE ACCREDITED AACSB ACCREDITED EQUIS ACCREDITED ASSOCIATION OF MBAS ACCREDITED

APPENDIX 2: Consent Form

Project Title:

An exploration of strategic decision-making around the adoption and implementation of AI technologies into organisations/working environments from the perspectives of key organisational stakeholders.

Lead researcher: Ali Gordjahanbeiglou

Please circle Yes or No

1	I confirm that I have read the Participant Information Sheet for the above study. I have had an opportunity to consider the information, ask questions and have had these answered satisfactorily.	Y	N
2	I understand that my participation is voluntary, and that I am free to withdraw from this research at any time up to data analysis stage, without giving any reason and without my legal rights being affected.	Y	N
3	I agree to my personal data and other related data, being collected during the research and processed as described in the Participant's information sheet.	Y	N
4	I agree to my interview being audio recoded.	Y	N
5	I agree that the researcher can use anonymised quotes made by me in their PhD thesis and any publications that may be produced from this research.	Y	N
6	I agree to anonymised data pertaining to this company being used by the research team for future research.	Y	N
7	I have the English language skills necessary to take part in an interview	Y	N
8	I agree to take part in this research.	Y	N

Please delete as appropriate:

- I would like a summary of the interview (via Email/ via post)
- I would like a report of the research's findings (via Email/ via post)

Preferred address / email for feedback, interview summary if applicable:

Name of participant: Date: Signature:

Researcher: Ali Gordjahanbeiglou Date:
Signature:

Ali Gordjahanbeiglou

APPENDIX 3: Project Brief

Aston Business School
BIRMINGHAM UK

Aston Business School
Aston University
Birmingham B4 7ET
United Kingdom

+44 (0)121 204 3000
aston.ac.uk/abs

Research Project Working Title:
An exploration of strategic decision-making around the adoption and implementation of AI technologies into organisational work practices.

Rationale and Background:
The breakthrough of Artificial Intelligence (AI) technologies has revolutionised the nature of business and organisations in the fourth industrial era, offering firms new and innovative opportunities in terms of data processing, decision-making and sophisticated solutions for work practices (e.g., Makarius et al., 2020). Yet, as with all new technologies, the emergence of AI has also raised important questions regarding its introduction, governance, effectiveness and impact on the working lives of employees (e.g., Abubakar et al., 2019). One particular avenue of academic enquiry seeks to better understand how and why decisions are made within organisations to introduce, design and implement AI, arguing that it is these decisions – and the decision-makers involved – that will influence the long-term impact and effectiveness of any new AI technology (e.g., Talamo et al., 2021). Our research, therefore, seeks to work with organisations who have recently introduced (or are considering introducing) AI into their work practices. This is to gain a better understanding in terms of the decision-making processes that have influenced the organisations' final decision to introduce AI, its nature and role it will play. We are particularly interested in the key factors considered when making decisions to introduce AI. As such, our main research questions are as follows:

Main Research Questions:
1. What are the drivers for introducing AI into work practices?
2. Who are the key stakeholders involved in this decision?
3. What factors are considered when designing and implementing AI technologies at work?
4. How effectively was AI introduced and implemented into the organisation?
5. How effective has the new AI been?

Aston Business School
BIRMINGHAM UK

Aston Business School
Aston University
Birmingham B4 7ET
United Kingdom

+44 (0)121 204 3000
aston.ac.uk/abs

Research Methods:
Our project follows a qualitative case-study research design.

Data Collection

i. One-to-one semi-structured interviews (approx. 30-35) with the following stakeholders:
   - Key decision-makers involved in AI introduction and implementation
   - Line managers of teams where AI is implemented
   - Employees working with AI
   - Other key stakeholders, where appropriate.

Of course, we would appreciate your guidance in identifying the relevant stakeholders involved in the decision to introduce AI into your own working practices, and potential individuals who may be willing to participate in this research. We would like to make clear that participation in interviews will be entirely voluntary and individuals will be free to withdraw at any stage of the process. In light of the global pandemic and social distancing safety measure, we have a duty of care to all participants and, as such, we have the option to conduct interviews virtually through Microsoft Teams or Zoom. The option of face-to-face meetings will be discussed with potential interviewees, and only with their permission will both options be offered to participants.

ii. In addition to the above, it would also be useful to arrange one or two onsite visits to familiarise the researcher with the organisation and AI. This procedure will include observations of AI-based systems and team members who are directly interacting with the AI. The researcher will ensure that his presence will not interfere with any on-going procedure or daily tasks of team-members.

We will provide the organisation with an overview of the research process and how it is managed in terms of ethical considerations, and a consent form for the potential participants. We will also share interview questions with all participants and your research team in advance of data collection commencing.

Data Analysis
Only with the participants' permission will interviews be recorded. Recordings will be transcribed and stored in Aston 'Box' cloud storage, which is a requirement of the Aston

Aston Business School
BIRMINGHAM UK

Aston Business School
Aston University
Birmingham B4 7ET
United Kingdom

+44 (0)121 204 3000
aston.ac.uk/abs

University Research Committee. This storage is secured with login credentials and is only accessible by the lead researcher. Data transcripts, and written notes for those who choose not to be recorded, will be analysed by Thematic Analysis and NVivo, which is a software package used by qualitative researchers. During this process, we will ensure that participants' names and identities are removed from the data.

Ethical Considerations

All stages of the research design, including data collection and analysis, are reviewed, and have been signed off by the Aston University Research Ethics Committee. We are happy to seek approval from any equivalent ethics committee upon request. Individual participation in our research will be voluntary and participants will be free to withdraw at any time. All participants will be anonymized, and all findings will be fed back in themes to maintain confidentiality. All data will be collected and stored securely in adherence to UK Data Protection laws, and 2018 GDPR regulations. All details about the research process will be available for potential participants in the participant information sheet. Clarification that participation is voluntary, participant rights for withdrawal and matters of privacy, confidentiality, anonymity, as well as the option for participants' interviews to be recorded or not, will also be explained in the participant consent form. Consent forms will be signed by participants prior the interview taking place.

Our research data collection process will be in accordance with Economic and Social Research Council-ESRC and GDPR guidelines.

Proposed Outcomes:

1. A written Evaluation Report outlining our key findings and recommendations. We would also be happy to present our findings to any individuals, committees or events that you feel would benefit. Our main aim is to provide your organisation with an opportunity to receive invaluable insight, and feedback, into the processes involved in deciding, designing and implementing AI technologies into your work practices. This Evaluation Report would deliver this insight and feedback and we would be keen to work with you to decide its exact nature, focus and content.
2. The data collected will also form the basis of my PhD Thesis, and at least one academic journal article and one practitioner-focussed output. Again, these would be made available to anyone who would be interested.

Aston Business School
BIRMINGHAM UK

Aston Business School
Aston University
Birmingham B4 7ET
United Kingdom

+44 (0)121 204 3000
aston.ac.uk/abs

References:

Abubakar, A. M., Behravesh, E., Rezapouraghdam, H., & Yildiz, S. B. (2019). Applying artificial intelligence technique to predict knowledge hiding behavior. *International Journal of Information Management*, 49, pp. 45–57.

Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, 120, pp. 262–273.

Talamo, A., Marocco, S. and Tricol, C. (2021) '"The Flow in the Funnel": Modeling Organizational and Individual Decision-Making for Designing Financial AI-Based Systems', *Frontiers in Psychology*, 12. doi: 10.3389/fpsyg.2021.697101.

This research is self-funded and is being supervised by academic members of Aston University Business School. The Research Team contact details are shared below. If you have any questions and wish further details, please do not hesitate to contact us.

Ali Gordjahanbeiglou: Tel: 07934 197 292. Email: gord1901@aston.ac.uk

Dr Jonathan R Crawshaw: Tel: 0121 204 3130. Email: j.r.crawshaw2@aston.ac.uk

Prof Nicholas Theodorakopoulos: Tel: 0 121 2043472. Email:

n.theodorakopoulos@aston.ac.uk

Dr Judy Scully: Tel: 0121 204 3229. Email : j.w.scully@aston.ac.uk

APPENDIX 4: Interview Schedule

A list of Interview Questions:

Management (Directors [includ. HR, etc.]) :

1. In your opinion, what are/were the key motivational factor for your organisation to adopt the Risk Assessment algorithm(s) (e.g., [redacted] etc.) in the first place?
2. How is (are) the Risk Assessment algorithm(s) making contributions towards the organisational processes?
3. How are strategic decisions (such as adoption of Risk Assessment algorithm) made in your organisation?
4. Who are the key organisational stakeholders that participate in the strategic decisions to introduce/implement Risk Assessment algorithms?
5. How is this stakeholder participation managed?
6. In what ways do HR practitioners influence in the decisions regarding the process of Risk Assessment algorithm(s)?
7. In what ways do HR /People Group contribute and support the employees in terms of well-being, training, reassurance, at the time of transformation (e.g., Introduction of Risk Assessment algorithm(s)?
8. Were HR issues discussed when deciding to adopt tools? Please explain.
9. Were ethical issues discussed when deciding to adopt the Risk Assessment algorithm(s)? Please explain.
10. How is the performance of the Risk Assessment algorithm(s) monitored and evaluated? Who are the responsible stakeholders in the process of monitoring and evaluation?
11. Have you implemented any measures/ protocols to oversee the decisions of Risk predicting algorithm(s)?
12. Who has designed, and monitors these protocols?
13. To what extent is the adoption and governance of Risk Assessment algorithm(s) in accordance with ethical regulation provided by related bodies (such as the Alan Turing Institute ethical frameworks)?
14. Overall, do you believe the adoption of Risk Assessment algorithm(s) has been a success? Why/why not?

List of Interview Questions:

IT Teams (Data Scientists, Digital support etc.):

1. Are there any on-going plans to adopt technologies based on AI/algorithms at [redacted] (and/or across its services [redacted])?
2. What are/were the key motivating factors for [redacted] to design and implement such systems?
3. How is (are) the AI/algorithmic solutions making contributions towards the organisational processes?
4. How are the strategic decisions (such as development of AI tools) made in [redacted]?
5. Who are the key organisational stakeholders that participate in the strategic decisions leading to design and implementation of AI agents?
6. How is the stakeholder participation managed?
7. In your opinion what are the HR/ people management issues in terms of design and implementation processes of AI or similar algorithmic tools?
8. In your opinion what are the Ethical implications and/or challenges of AI/algorithms?
9. Were Ethical issues discussed in the process of design and implementation of these technologies? Please explain.
10. How is the performance of implemented AI tools monitored and evaluated? Who are the responsible stakeholders in the process of assessment and evaluation?
11. To what extent are the design and governance of your AI systems in accordance with the frameworks offered by related bodies (such as [redacted] -by *the Alan Turing institute*)?
12. Overall, do you believe adoption and implementation of AI solutions have been successful? Why/why not?

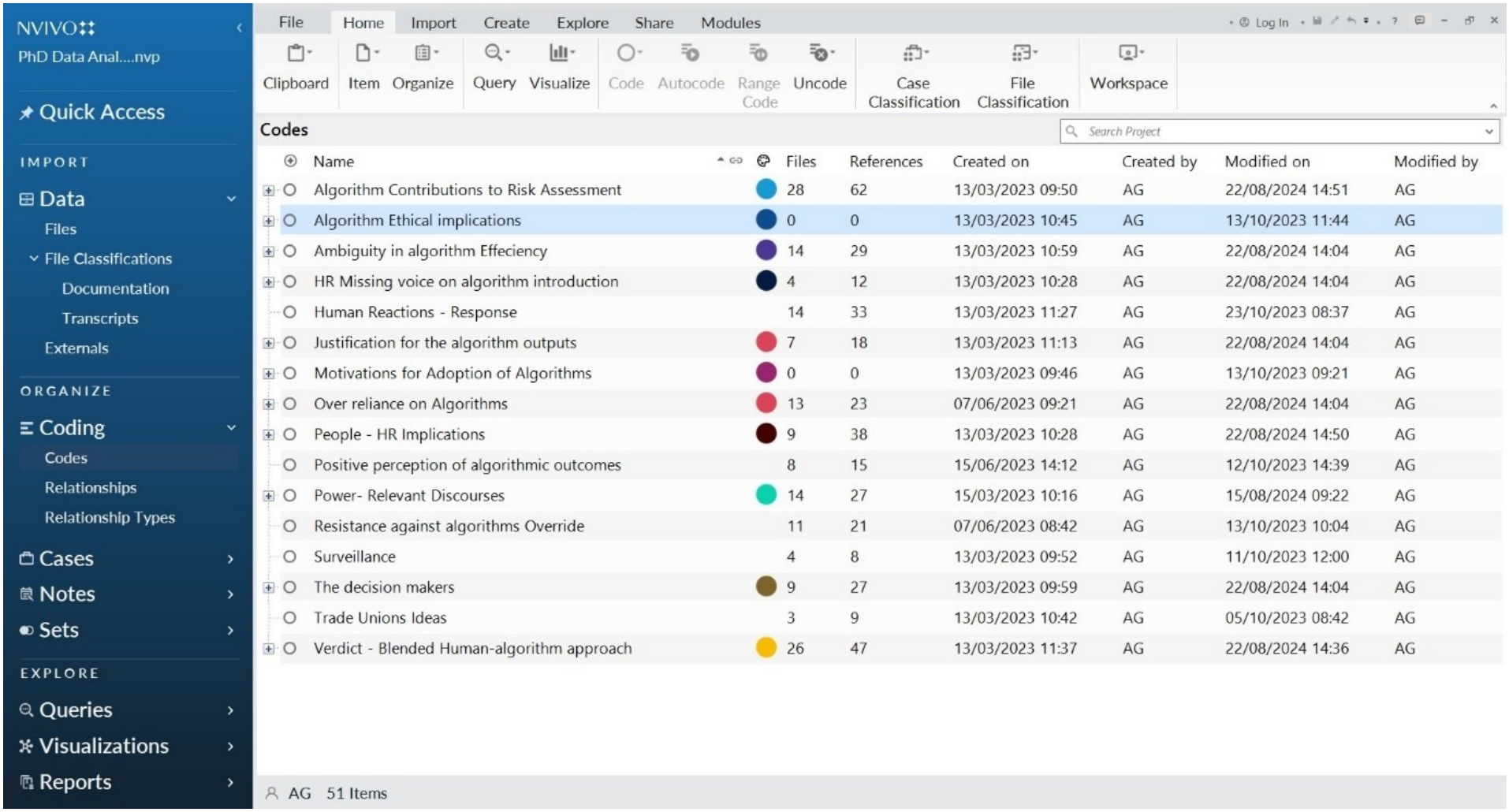
HR /OD Practitioners:

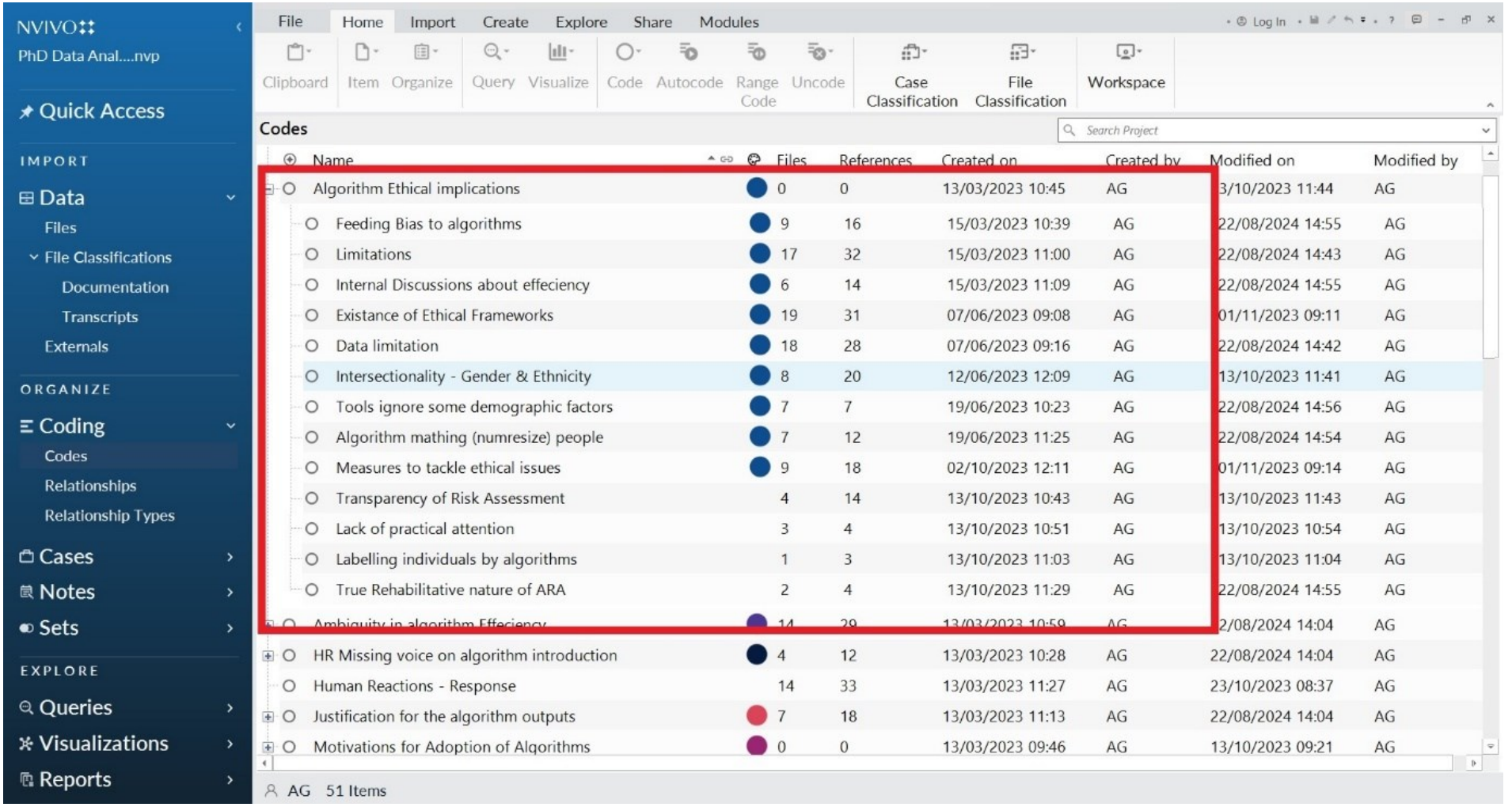
1. In your opinion, what are/were the key motivational factor for [redacted] to adopt Algorithmic tools (the [redacted]) for Professional assessments?
2. Who are the key organisational stakeholders that participate in the strategic decisions to introduce/implement the Predictive algorithm?
3. How involved were HR in these decisions? What was the nature of this involvement?
4. Were you happy with level of influence and input HR had in these decisions? Why/why not?
5. In your opinion, what are the HR/employee challenges in terms of Human-Algorithm interactions?
6. Were HR issues discussed when deciding to adopt the Predictive algorithm? Please explain.
7. Were any ethical implications discussed when making decisions to implement the Predictive algorithm? Please explain.
8. In what ways does the HR department contribute and support the employees in terms of well-being, training, reassurance, with respect to work transformations due to the Predictive algorithm?
9. How is the performance of the algorithm monitored and evaluated? Who are the responsible stakeholders in the process of monitoring and evaluation?
10. Have you implemented any measures/ protocols to oversee the outputs of the Predictive algorithm?
11. Who has designed, and monitors these protocols?
12. To what extent is the adoption and governance of the Predictive algorithm in accordance with ethical regulation provided by related bodies (such as [redacted]- The Alan Turning Institute, IEEE or AI4People)?
13. Overall, do you believe the adoption of the Predictive algorithm has been a success? Why/why not?

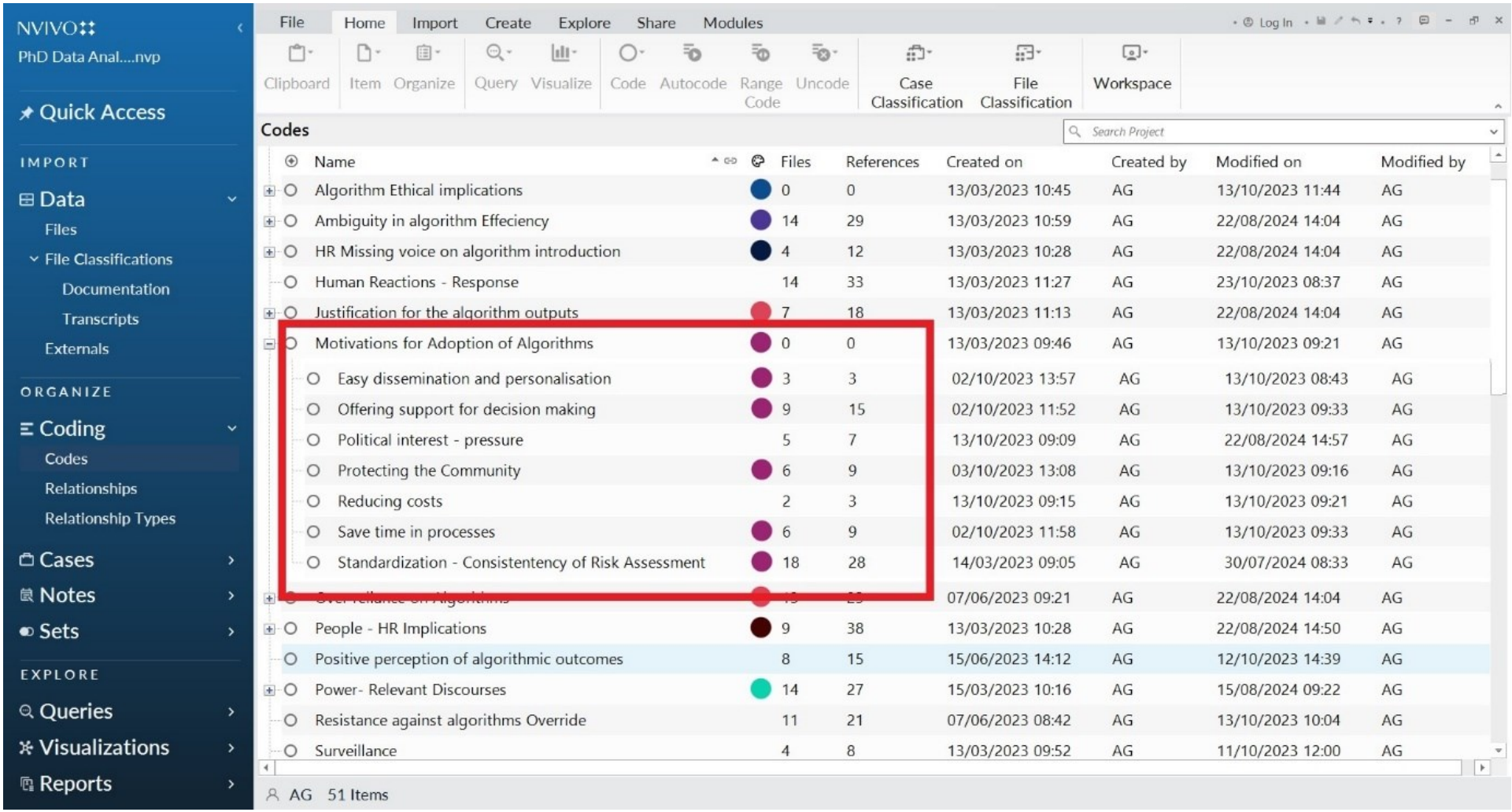
(Probation Officers- Seniors; Line managers, etc.):

1. Do you work/ interact with the Algorithmic tools, e.g., [redacted]? Please explain.
2. What were your feelings / reactions when you were informed that your work now includes an algorithm?
3. Do you have any concerns about the about the adoption of the Risk Assessment algorithms?
4. Are you able to express these concerns, and are they listened to?
5. Do you feel you have had the support (in terms of trainings, well-being initiatives etc.) for the work transitions due to Risk Assessment algorithms? Please explain.
6. Has HR provided you appropriate support since the adoption of the Risk Assessment algorithm?
7. In your opinion, what are/were the key motivational factor for your organisation to adopt the Risk Assessment algorithm(s) in the first place?
8. Who were the key organisational stakeholders that participate in the strategic decisions to introduce/implement the Risk Assessment algorithm(s)?
9. How involved were HR in these decisions? What was the nature of this involvement?
10. Did you feel you that you had an influence on this decision? Please explain.
11. Were HR issues discussed when deciding to adopt the Predictive algorithm(s)? Please explain.
12. Were ethical implications of AI adoption discussed when making decisions to implement the Risk Assessment algorithm(s)? Please explain.
13. Are you aware of your organisation's ([redacted] particularly) ethical guidelines or code of conduct relating to AI / Algorithms?
14. How is the performance of the Risk Assessment algorithm(s) monitored and evaluated? Who are the responsible stakeholders in the process of monitoring and evaluation?
15. Overall, do you believe the adoption of the Risk Assessment algorithm(s) has been a success? Why/why not?

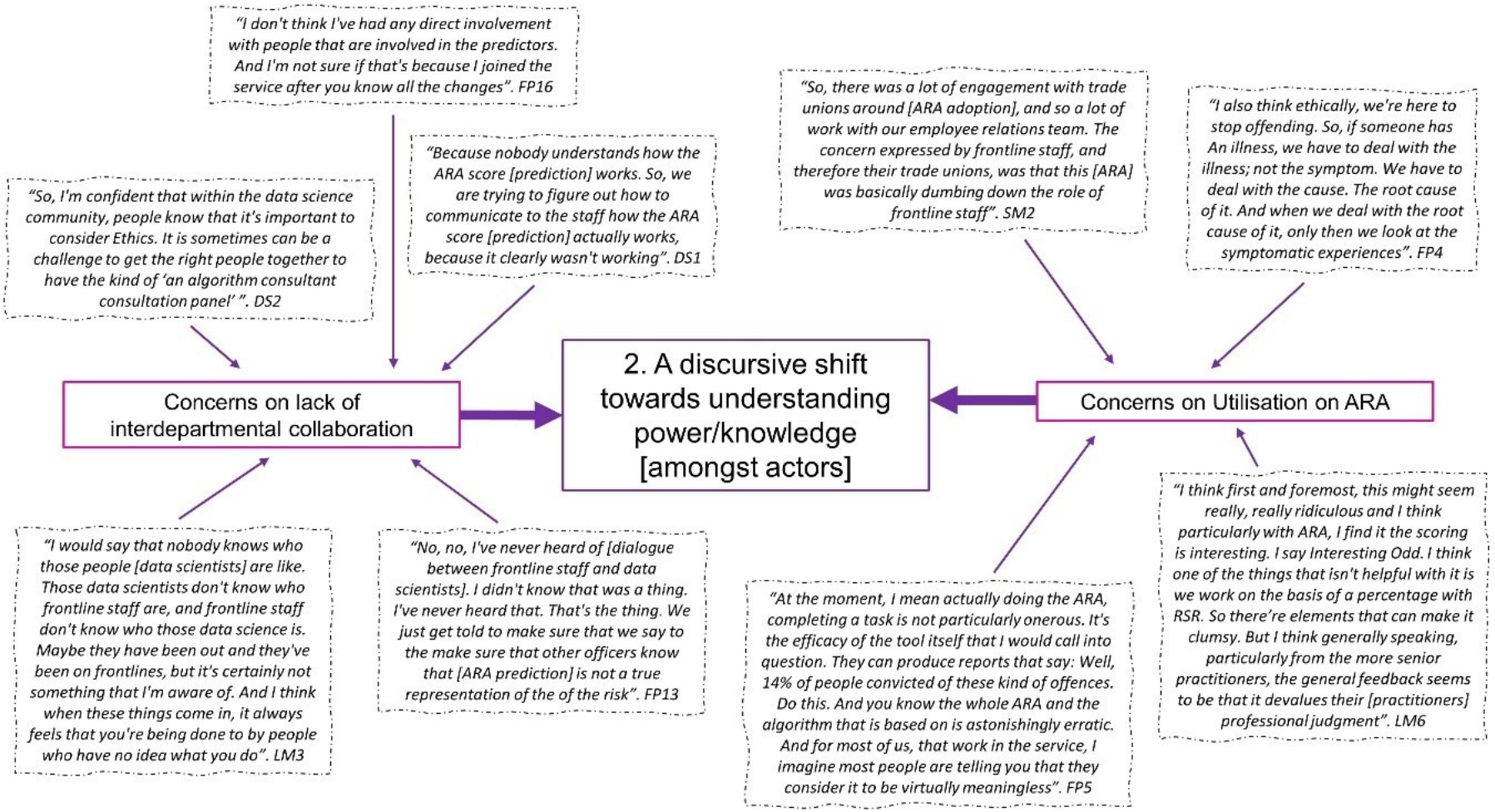
APPENDIX 5: Screenshots of coding

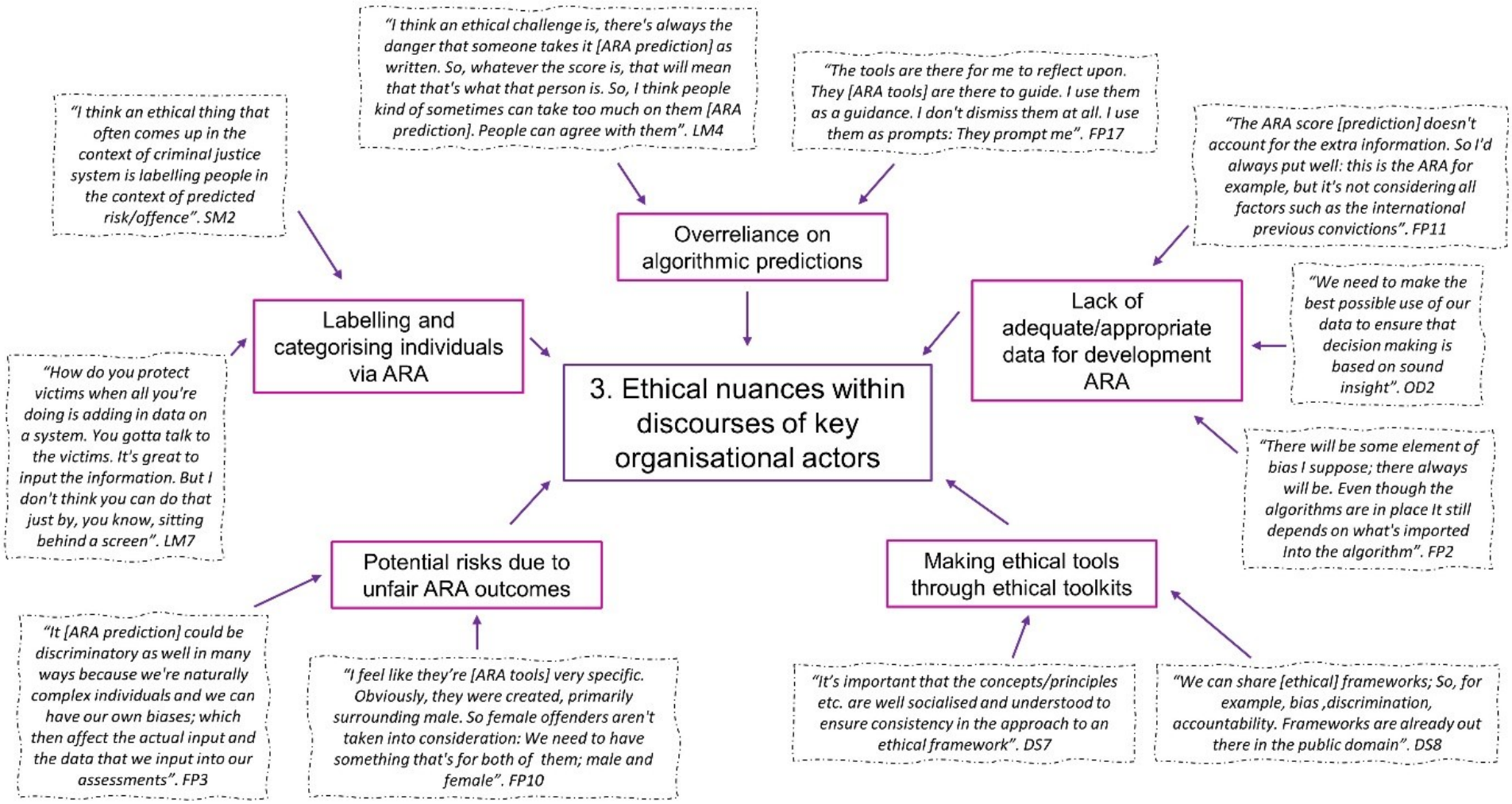


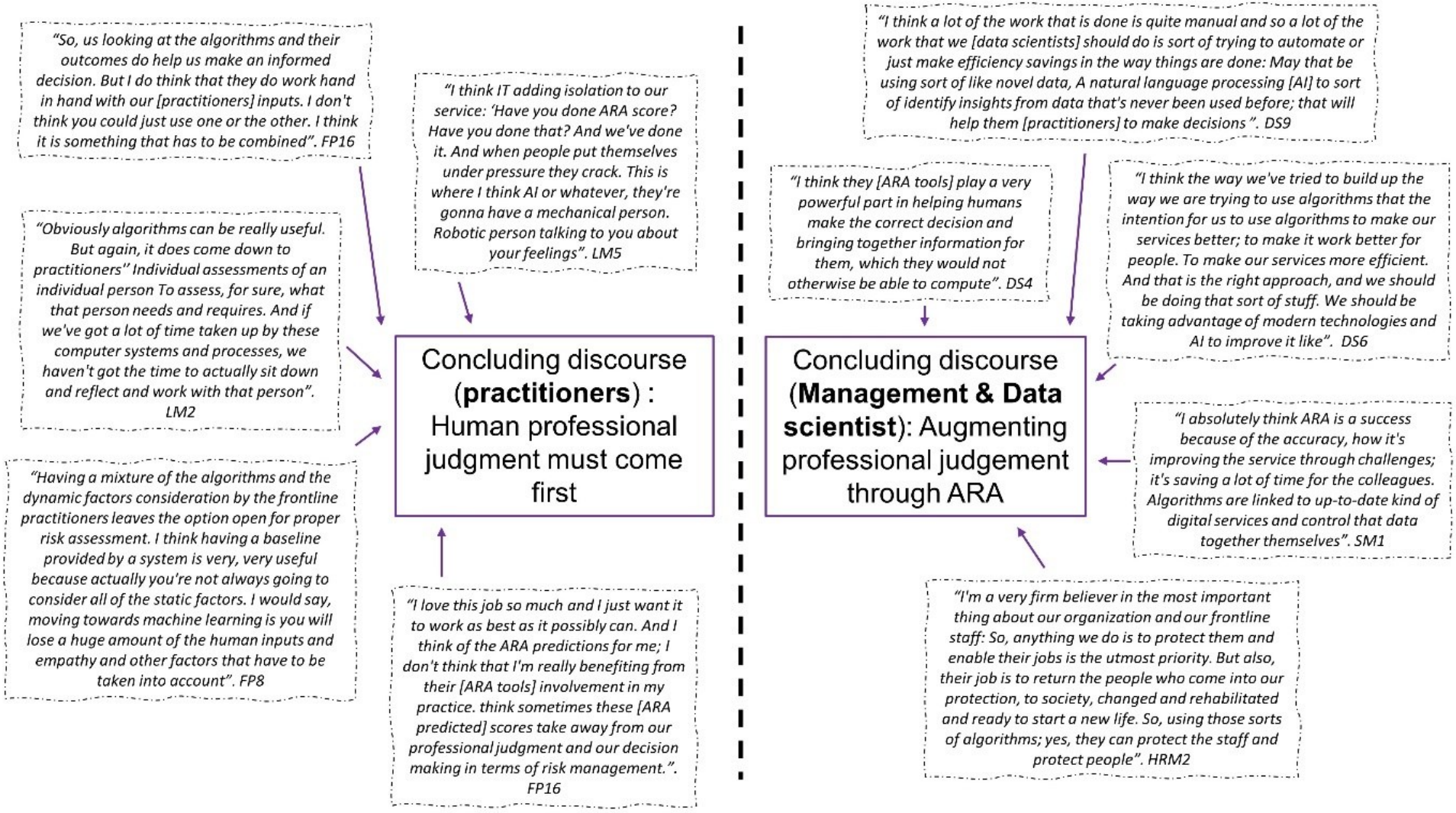

NVIVO
PhD Data Anal....nvp
Quick Access
IMPORT
Data
Files
File Classifications
Documentation
Transcripts
Externals
ORGANIZE
Coding
Codes
Relationships
Relationship Types
Cases
Notes
Sets
EXPLORE
Queries
Visualizations
Reports
File
Home
Import
Create
Explore
Share
Modules
Log In
Clipboard
Item
Organize
Query
Visualize
Code
Autocode
Range Code
Uncode
Case Classification
File Classification
Workspace
Codes
Search Project
Name
Files
References
Created on
Created by
Modified on
Modified by
Algorithm Ethical implications 0 0 13/03/2023 10:45 AG 3/10/2023 11:44 AG
Feeding Bias to algorithms 9 16 15/03/2023 10:39 AG 22/08/2024 14:55 AG
Limitations 17 32 15/03/2023 11:00 AG 22/08/2024 14:43 AG
Internal Discussions about effeciency 6 14 15/03/2023 11:09 AG 22/08/2024 14:55 AG
Existance of Ethical Frameworks 19 31 07/06/2023 09:08 AG 01/11/2023 09:11 AG
Data limitation 18 28 07/06/2023 09:16 AG 22/08/2024 14:42 AG
Intersectionality - Gender & Ethnicity 8 20 12/06/2023 12:09 AG 13/10/2023 11:41 AG
Tools ignore some demographic factors 7 7 19/06/2023 10:23 AG 22/08/2024 14:56 AG
Algorithm mathing (numresize) people 7 12 19/06/2023 11:25 AG 22/08/2024 14:54 AG
Measures to tackle ethical issues 9 18 02/10/2023 12:11 AG 01/11/2023 09:14 AG
Transparency of Risk Assessment 4 14 13/10/2023 10:43 AG 13/10/2023 11:43 AG
Lack of practical attention 3 4 13/10/2023 10:51 AG 13/10/2023 10:54 AG
Labelling individuals by algorithms 1 3 13/10/2023 11:03 AG 13/10/2023 11:04 AG
True Rehabilitative nature of ARA 2 4 13/10/2023 11:29 AG 22/08/2024 14:55 AG
Ambiguity in algorithm Effeciency 14 29 13/03/2023 10:59 AG 2/08/2024 14:04 AG
HR Missing voice on algorithm introduction 4 12 13/03/2023 10:28 AG 22/08/2024 14:04 AG
Human Reactions - Response 14 33 13/03/2023 11:27 AG 23/10/2023 08:37 AG
Justification for the algorithm outputs 7 18 13/03/2023 11:13 AG 22/08/2024 14:04 AG
Motivations for Adoption of Algorithms 0 0 13/03/2023 09:46 AG 13/10/2023 09:21 AG
AG 51 Items


NVIVO
PhD Data Anal....nvp
Quick Access
IMPORT
Data
Files
File Classifications
Documentation
Transcripts
Externals
ORGANIZE
Coding
Codes
Relationships
Relationship Types
Cases
Notes
Sets
EXPLORE
Queries
Visualizations
Reports
File
Home
Import
Create
Explore
Share
Modules
Log In
Clipboard
Item
Organize
Query
Visualize
Code
Autocode
Range Code
Uncode
Case Classification
File Classification
Workspace
Codes
Search Project
Name
Files
References
Created on
Created by
Modified on
Modified by
Algorithm Ethical implications 0 0 13/03/2023 10:45 AG 13/10/2023 11:44 AG
Ambiguity in algorithm Effeciency 14 29 13/03/2023 10:59 AG 22/08/2024 14:04 AG
HR Missing voice on algorithm introduction 4 12 13/03/2023 10:28 AG 22/08/2024 14:04 AG
Human Reactions - Response 14 33 13/03/2023 11:27 AG 23/10/2023 08:37 AG
Justification for the algorithm outputs 7 18 13/03/2023 11:13 AG 22/08/2024 14:04 AG
Motivations for Adoption of Algorithms 0 0 13/03/2023 09:46 AG 13/10/2023 09:21 AG
Easy dissemination and personalisation 3 3 02/10/2023 13:57 AG 13/10/2023 08:43 AG
Offering support for decision making 9 15 02/10/2023 11:52 AG 13/10/2023 09:33 AG
Political interest - pressure 5 7 13/10/2023 09:09 AG 22/08/2024 14:57 AG
Protecting the Community 6 9 03/10/2023 13:08 AG 13/10/2023 09:16 AG
Reducing costs 2 3 13/10/2023 09:15 AG 13/10/2023 09:21 AG
Save time in processes 6 9 02/10/2023 11:58 AG 13/10/2023 09:33 AG
Standardization - Consistentency of Risk Assessment 18 28 14/03/2023 09:05 AG 30/07/2024 08:33 AG
07/06/2023 09:21 AG 22/08/2024 14:04 AG
People - HR Implications 9 38 13/03/2023 10:28 AG 22/08/2024 14:50 AG
Positive perception of algorithmic outcomes 8 15 15/06/2023 14:12 AG 12/10/2023 14:39 AG
Power- Relevant Discourses 14 27 15/03/2023 10:16 AG 15/08/2024 09:22 AG
Resistance against algorithms Override 11 21 07/06/2023 08:42 AG 13/10/2023 10:04 AG
Surveillance 4 8 13/03/2023 09:52 AG 11/10/2023 12:00 AG
AG 51 Items

APPENDIX 6: Illustrations of reflexive thematic analysis + FDA




"I think an ethical challenge is, there's always the danger that someone takes it [ARA prediction] as written. So, whatever the score is, that will mean that that's what that person is. So, I think people kind of sometimes can take too much on them [ARA prediction]. People can agree with them". LM4
"The tools are there for me to reflect upon. They [ARA tools] are there to guide. I use them as a guidance. I don't dismiss them at all. I use them as prompts: They prompt me". FP17
"I think an ethical thing that often comes up in the context of criminal justice system is labelling people in the context of predicted risk/offence". SM2
"The ARA score [prediction] doesn't account for the extra information. So I'd always put well: this is the ARA for example, but it's not considering all factors such as the international previous convictions". FP11
Overreliance on algorithmic predictions
Labelling and categorising individuals via ARA
Lack of adequate/appropriate data for development ARA
"We need to make the best possible use of our data to ensure that decision making is based on sound insight". OD2
"How do you protect victims when all you're doing is adding in data on a system. You gotta talk to the victims. It's great to input the information. But I don't think you can do that just by, you know, sitting behind a screen". LM7
3. Ethical nuances within discourses of key organisational actors
"There will be some element of bias I suppose; there always will be. Even though the algorithms are in place It still depends on what's imported Into the algorithm". FP2
Potential risks due to unfair ARA outcomes
Making ethical tools through ethical toolkits
"It [ARA prediction] could be discriminatory as well in many ways because we're naturally complex individuals and we can have our own biases; which then affect the actual input and the data that we input into our assessments". FP3
"I feel like they're [ARA tools] very specific. Obviously, they were created, primarily surrounding male. So female offenders aren't taken into consideration: We need to have something that's for both of them; male and female". FP10
"It's important that the concepts/principles etc. are well socialised and understood to ensure consistency in the approach to an ethical framework". DS7
"We can share [ethical] frameworks; So, for example, bias ,discrimination, accountability. Frameworks are already out there in the public domain". DS8


"So, us looking at the algorithms and their outcomes do help us make an informed decision. But I do think that they do work hand in hand with our [practitioners] inputs. I don't think you could just use one or the other. I think it is something that has to be combined". FP16
"I think IT adding isolation to our service: 'Have you done ARA score? Have you done that? And we've done it. And when people put themselves under pressure they crack. This is where I think AI or whatever, they're gonna have a mechanical person. Robotic person talking to you about your feelings". LM5
"Obviously algorithms can be really useful. But again, it does come down to practitioners" Individual assessments of an individual person To assess, for sure, what that person needs and requires. And if we've got a lot of time taken up by these computer systems and processes, we haven't got the time to actually sit down and reflect and work with that person". LM2
"Having a mixture of the algorithms and the dynamic factors consideration by the frontline practitioners leaves the option open for proper risk assessment. I think having a baseline provided by a system is very, very useful because actually you're not always going to consider all of the static factors. I would say, moving towards machine learning is you will lose a huge amount of the human inputs and empathy and other factors that have to be taken into account". FP8
Concluding discourse (practitioners) : Human professional judgment must come first
"I love this job so much and I just want it to work as best as it possibly can. And I think of the ARA predictions for me; I don't think that I'm really benefiting from their [ARA tools] involvement in my practice. think sometimes these [ARA predicted] scores take away from our professional judgment and our decision making in terms of risk management.". FP16
"I think a lot of the work that is done is quite manual and so a lot of the work that we [data scientists] should do is sort of trying to automate or just make efficiency savings in the way things are done: May that be using sort of like novel data, A natural language processing [AI] to sort of identify insights from data that's never been used before; that will help them [practitioners] to make decisions ". DS9
"I think they [ARA tools] play a very powerful part in helping humans make the correct decision and bringing together information for them, which they would not otherwise be able to compute". DS4
"I think the way we've tried to build up the way we are trying to use algorithms that the intention for us to use algorithms to make our services better; to make it work better for people. To make our services more efficient. And that is the right approach, and we should be doing that sort of stuff. We should be taking advantage of modern technologies and AI to improve it like". DS6
Concluding discourse (Management & Data scientist): Augmenting professional judgement through ARA
"I absolutely think ARA is a success because of the accuracy, how it's improving the service through challenges; it's saving a lot of time for the colleagues. Algorithms are linked to up-to-date kind of digital services and control that data together themselves". SM1
"I'm a very firm believer in the most important thing about our organization and our frontline staff: So, anything we do is to protect them and enable their jobs is the utmost priority. But also, their job is to return the people who come into our protection, to society, changed and rehabilitated and ready to start a new life. So, using those sorts of algorithms; yes, they can protect the staff and protect people". HRM2

APPENDIX 7: A sample of generated auto transcripts (Teams)

0:0:0.0 --> 0:0:6.470
Ali Gordjahanbeiglou (Research Student)
I think it's it's more appropriate if I introduce myself a little bit. So my name is Ali Ali gorjan beglau.
It's a bit difficult surname, but please feel free to call me Ali. I'm a second yeah, pH D candidate at Aston Business Score Department of Work and organization. So My day to day life at the moment is basically collecting data, attending conferences, field work.

And. So it's great to have you here today. So you you have my gratitude and yeah, please, please over to you.

0:0:36.110 --> 0:1:6.10
Data Scientist 1
OK, so I'm a data scientist, but from a uh, unusually from a data scientist, I'm from a kind of Social Research background, so I've got academic background variously in demography, statistics and forensic psychology research. My specialist subject, if you like, is offender assessment, particularly actuarial assessment of Reoffending risks. I've been working in the last couple of decades on topics yeah related to the offender assessment system that we use in the [redacted] and the actuarial systems Related to that, essentially. so although I'm as I'm in government, I don't publish all that much, but there are a few publications of mine out there which may be helpful though concerning my citation list afterwards, which might help just to chase some things up that I might mention.

0:1:45.780 --> 0:2:1.880
Ali Gordjahanbeiglou (Research Student)
That's great. Thank you so much. Uh, so I mean, I mean, I understand that in our communications you mentioned they're not particularly artificial intelligence technologies or algorithms, but they're mainly tools, predictive tools, so.

0:2:2.530 --> 0:2:16.200
Data Scientist 1
Yes. So that I say they're algorithms, but they're not very much artificial intelligence. You know, the decisions were very much made, you know, by me as the modeler and Yeah, the essentially the, you know, the computer was working as an advanced calculator and well, you know there wasn't sort of automated detection of effects and interactions and so forth like you might get with tree based systems or neural network systems or anything like that, yeah.

0:2:36.890 --> 0:2:49.40
Ali Gordjahanbeiglou (Research Student)
That's great, but just going to the past to the I think it was 2014, but [redacted] told me that you introduced this technology, so I wanted to ask for the first question. What were Motivating factors [redacted] to actually adopt and implement these technologies.

0:2:58.180 --> 0:3:1.180
Data Scientist 1
Well, OK. Um now.

0:3:0.460 --> 0:3:2.10
Ali Gordjahanbeiglou (Research Student)
The the reasons basically.

0:3:2.320 --> 0:3:13.900
Data Scientist 1
Yes, now some of this will be about the workings of government and stuff that's not been fully published. So I mean I, you know, Jim has given permission for this interview.
But it's probably the kind of thing that we can repeat verbatim.
Still.
Umm there was.
Yeah. OK. So it's, you know, it's a matter of public record that the probation service was broken up. As a matter of government policy.
Umm it was.
The desire was to reduce the cost of running the [redacted]. Um, it was also hoped that innovation would uh result From having a different [redacted] run by uh, different organizations. UM.
And.
But there were concerns that actually if you had a marketization of the full probation caseload with every offender being managed by either private sector organizations or maybe consortia of charities with private sector partners, that ended up happening, that they might not be skilled at managing the most dangerous offenders. So therefore a system was wanted. For deciding who are the most dangerous offenders, who should be reserved uh to be managed by a rump [redacted] that was run in the public sector with some of the most experienced staff and so forth.
And so.
Uh, we knew from previous work of mine that The professional judgments, the judgments that probation officers make about who is the most dangerous, uh can be inconsistent.
And there's also an issue that it would be difficult that you would want to allocate the cases between the National Probation Service and these marketized organizations. The community rehabilitation companies. They were called, you would need to make that decision at the point of sentencing and therefore actually a full professional assessment could not always be done. And so therefore, you go to an actuarial assessment.

So for that reason, we devised this, um, risk of serious recidivism actuarial tool. So we already had some actual tools and use, but this tool the RSR had as the criterion that somebody would be convicted reconvicted of a new, serious offense. We went for a process of agreeing what should count as a serious offense. But you know lots of behaviours that potentially could.
Lots of different ways to do it, but we went for a structured process to make a decision about that.
Umm and then I used data from past [redacted] caseloads, followed them forward over time and modeled. What were these factors that were predictive of serious reoffending. now that?
The RSR score wasn't the sole factor in the process.
There was also whether offenders were managed under Mapper MAPPA, which is multi agency

public protection arrangements and that requires liaison between the probation organization and the police and various other agencies like social services in many cases.
Umm. And that was felt to be too complex and difficult in terms of statutory duties. Uh to have marketed out so that those cases state with the [redacted] as well.

And in the end, this is the bit that might be difficult to write. The ministers were concerned about it being a solely actuarial system because the macro rules are accurate and so, you know, the macro rules are an algorithm as well. They're not a mathematical algorithm, but if it's if someone was convicted of this offense within this time scale and so forth, then they are a mapper case. There's a few discretionary mapper cases, but mostly it's mechanistic.

But yeah, their ministers got worried about leaving this the lack of discretion for people who might present this dangerous. So actually, if a case was professionally rated as high or very high risk of serious harm, then they will reserve for the [redacted] as well. So you have the Mapper cases, the cases of high RSR scores, and the cases that were high or very high risk of serious harm. And there were some smaller groups as well, such as people who might present a counterterrorism risk, but they were numerically insufficient.

? Yeah. Insignificant. The three main groups were those groups I mentioned. And of course they overlap. Many people were in more than one group.
But that's how it was done. And so there, you know, we did some modeling of how many people would be included in the [redacted] case load if you set the rules up in various different ways. And that's how. Yeah, that's how you came to have this ecosystem where the RSR score Coexisted, of course, various other criteria that we used in case allocation.
Yeah. So that's how it started.

0:8:30.220 --> 0:8:56.10
Ali Gordjahanbeiglou (Research Student)
Thanks so much. That was a very, very inclusive answer, but now I think you briefly mentioned the second answer to the second question, but it's the second question actually related to the technology to this systems it themselves. So I want to ask how they're contributing towards the offending or monitoring process, how are they changing the status quo? In a nutshell, you know your opinion.
How do you how do you analyze this?

0:9:0.20 --> 0:9:2.110
Data Scientist 1
Yeah. OK. So.
The RSR score Is designed to be informative and UM You know, when an offender gets an RSR score, that literally does represent our best estimate of their likelihood of You know, committing a serious offense and getting convicted for it and uh, uh, yeah, have that serious reoffending outcome. And that is intended to help Probation practitioners, probation officers, basically to make decisions about how they manage their offenders and who they are, exercise most caution with about their risk, their risk management and so forth. So the idea is that they look at the offenders RSR score and then they consider other factors that they observe as well and use their experience and what they think of the offenders motivation, what they think of their living situation, their patterns of behaviour and so forth. Combine all that with the RSR score and therefore come to their risk of serious harm rating.

There have been some pieces of work I've done that look at um, how much the risk of the risk of harm, serious harm ratings. These professional ratings are influenced by the RSR score.
Those piece of work, however, are not published.

So I can tell you about them. I would have to ask you to not refer to them.

0:10:42.700 --> 0:10:43.600
Ali Gordjahanbeiglou (Research Student)
I see. OK.

0:10:44.240 --> 0:10:44.610
Ali Gordjahanbeiglou (Research Student)
Alright.

0:10:42.950 --> 0:10:47.400
Data Scientist 1
In what you write is that, uh, so if that's OK, I can go ahead, yeah.

0:10:46.420 --> 0:10:49.570
Ali Gordjahanbeiglou (Research Student)
Yeah, yeah, absolutely. Off the record. Yeah, yeah.

0:10:50.760 --> 0:10:51.250
Ali Gordjahanbeiglou (Research Student)
Absolutely.

0:10:48.470 --> 0:10:53.300
Data Scientist 1
Yeah, yeah. So basically.

0:10:55.750 --> 0:11:6.670
Data Scientist 1
OK. So something that is published in past compendia is a comparison of the predictive validity of the professional risk of serious harm ratings and these act and actuarial ways of predicting reoffending. A predecessor of the RSR scoring fact, and what they show is that the risk of serious harm ratings. I'm gonna call it ROSH, we call it ROSH risk of serious harms of mouthful. So when I say Rosh, it's the professional judgment of risk of serious harm.
The rush ratings are poor predictors of who will seriously reoffend.
Whereas the actuarial scores are good predictors, um and so, you know, one could form an argument that one shouldn't use the professional judgment at all.

0:11:50.920 --> 0:11:51.160
Ali Gordjahanbeiglou (Research Student)
OK.

What isn't published are actually no something I can refer you to is some data from back between 2005 and 2008. So long time ago. But the Offender Assessment system, the ROSH system were basically the same. It was just before RSR came in. Looking at the consistency of ROSH ratings across different parts of the country because we can say we can get the offenders

record and say where they were in the country and there's different regional [redacted] that have.
You have changed over time, but the principle of basically the same, you know, national probation service is made up of lots of little local areas and they've got a management structure and it goes upwards you.

Yeah, huge amounts of variation in the way that risk of serious harm ratings are made across the country, basically. So somebody who is high risk of serious harm in one region might well be medium risk of serious harm in other regions.

If you take that through to what was being done from 2014 onwards, then that means that you would get offenders were identical profiles and such an offender would probably be managed by the [redacted] as a high risk case in various parts of the country and by the Community rehabilitation companies as a medium worst case in other parts of the country.

Yeah. So that's not good. You wanna have consistent practice. You know, it's difficult to say what the right answer is categorically on how you should do this, but having a system where people, you know where it's a post code lottery plainly not a good thing.

Back to for firmly off the record. There's some stuff that I'm actually doing right now. We're gonna do some sort of prototyping in the field to try and, like, do some visualizations of how the RSR score works, because nobody understands how the RSR score works, that's bit clear. It's been user.

There's been using research done.
It's clear that because of the association of the RSR score with this allocation system, MPs CRC.
There's practitioners out there who think that the RSR score is like an abstract number that was just that doesn't really mean anything that was just come up with the case allocation, and now that we've finished the system with the CRC's, now that the CRC's been abolished, it's [redacted]

It's a national probation service for everybody. Again, there are. Why do we have to keep doing this stupid RSR score? What's the point? They haven't read the risk of serious harm guidance, which has been out there for two years, which tells them start with the RSR score. Add in your observations.
End up with the ROSH.

And yeah, and I'm looking at data from 2019 uh, and doing some visualizations of it just now. That shows that in South London, people were far less likely to be rated as high ROSH than in north London I broke down the basically the in London, the probation organization is more or less by London Borough. And so you can compare different London boroughs and say, OK, in Bromley and in Camden, they were both supposed to have about 10% of people be high risk of

serious on, but actually in Bromley it was 6% and in Camden it was 14%, that kind of thing. That degree of difference?

So we are trying to figure out how to communicate to the staff how the RSR score actually works, because it's clearly it wasn't working.

Why wasn't it working? Why people don't know so much? There is obviously responsibility on us as advisors of the tool, and there's also the fact that in 2014, apparently there are about 600 different sort of probation instructions and communications about the massive system change that was going on at the time. So yeah, we did put communications out there.

But it's not surprising that not many people read them. Uh, and of course, people really worried about, you know, their jobs and who they're gonna be working for, whether they're gonna get sacked.

And eventually some them did get sacked. More experienced staff who went to the CRC's, they tended to a year or two later, have basically got sacked cause they are more expensive than they had cheaper people brought in. So yeah, so no wonder that the communication didn't really work in those days.

Umm. And so, yeah. And evidently the training function hasn't been good enough on that. So we're now trying to play catch up. And after all of these years of people using RSR trying to find a way of educating them for what it actually is and how it can actually help them so.

Uh, yeah, I've realized it's frustrating that a lot of this is not published and probably a lot of it is stuff that Jim wouldn't want to appear in the public record because it is an interesting example of how you can have an actuarial tool that's out there.

We did validate it. Initially we're doing another validation that we published should think by the end of this year.

That it's natural that does work. It's in there. People are using it, but they haven't been told what it does. They misunderstand what it does. They hate it because they it. They think it's Like insulting their professional judgment and all of that.

0:16:54.300 --> 0:17:1.710
Ali Gordjahanbeiglou (Research Student)
Great. I would. I will make sure that this does few minutes of your remarks will not be included and or any reference to those.

0:17:0.180 --> 0:17:12.470
Data Scientist 1
Yeah, I'm. I mean, I think you could if you want to sort of write up for your own notes and then if you're circulating the note, then you could say [redacted], you know, what are you happy to

include in the public?
Uh, the you know, potentially including the public domain, what would you like me to redact so we could give Jim that opportunity?

0:17:19.130 --> 0:17:23.910
Ali Gordjahanbeiglou (Research Student)
You know, we'll get back to you on that. Maybe before it was being published, yeah.

0:17:31.530 --> 0:17:31.870
Ali Gordjahanbeiglou (Research Student)
OK.

0:17:20.970 --> 0:17:35.160
Data Scientist 1
Yeah. Yeah, yeah, yeah, absolutely. So, you know, I think so long as your note says, you know, Phillip indicated that some of this might not be fit for the public, but. And then Jim will be reassured. But. Yeah, yeah. Then then he can make a choice. Yes. OK.

0:17:45.840 --> 0:17:46.120
Data Scientist 1
Umm.

0:17:33.680 --> 0:17:52.850
Ali Gordjahanbeiglou (Research Student)
There's definitely. Absolutely thank you. So considering the adoption or implementation of this predictive tools in business world and organisational world as a strategic decision that is in the first question, you briefly mentioned that it's saving a lot of money and it's speeding up the processes so.

0:17:53.740 --> 0:18:10.90
Ali Gordjahanbeiglou (Research Student)
I want to I want to dissect how these strategic decisions as such adoption of this predictive tools are made initially, how, I mean I understand the idea came from but how this decision was made and?

0:18:10.680 --> 0:18:11.310
Data Scientist 1
Yeah.

0:18:11.280 --> 0:18:15.250
Ali Gordjahanbeiglou (Research Student)
Who are those decision makers who were brainstorming in that sense?

0:18:16.740 --> 0:18:20.490
Data Scientist 1
OK, so the key thing is that is, yeah, we're in government.

And so the decision to break up the probation service and have these community rehabilitation companies funded the fundamentally came from the government ministers. That was the strategic direction that they demanded. Umm. And so [redacted]. Uh, was the Secretary of State uh for justice. The Lord again. AKA T[redacted] that time.

He is somewhat notorious amongst people who have been government ministers. Uh, you can look up his record and make your own judgment.

But it was very much his idea that we would do this. So the idea basically was that running [redacted] in this way would save money and hopefully would lead to innovation.

An uncontroversially good thing that he wished to achieve, and that this money was designed to do cause. Of course we were under the, you know, the sort of austerity, years of the, you know, 2010 to 2015 government here. So proposals have to be cost neutral. But until this came in, if someone was released from prison.

And they're prison sentence had been under one year, then they would come out of prison and go into the community, and they would have no support from probation.

So now everybody has a year of support when they come out of prison, so long as they're prison sentence. Them was at least two days. So if you get a two day prison sentence now, you'll spend one day in prison one day on license and then 364 bonus days of post sentence supervision that only exist thanks to these changes. So that's what this was intended to finance.

In fact, the CRC's did not make a saving they cost more, they cost more money, they the providers weren't able to run them as cheaply as they expected, and they went cap in hand for bailouts. And eventually you know the there's a lot of criticism. Eventually the whole thing was you know, abolished it as of the middle of last year, we have the National Unified Probation Service. Good. But that the you asked me about the decision decision came from ministers to do this. That was the strategy. So then you have to get to the tactics and basically we're talking about people working out options here.

So we were given the brief of devising an actuarial system to predict, you know, identify the most dangerous offenders. So we said, all right then, if they've already done some exploratory work about predicting homicide and wounding and about predicting sexual reoffending.

Let's take that. Let's come up with a definition of serious reoffending and build a predictor that will work for all sorts of serious reoffending. So we talked to lots of we checked out what is sort of already in use when people are looking at serious offenses like there's a thing where called serious further offences, where if someone on probation and they get convicted, one of these offenses, they're sort of automatically a review of the case. Did we do things wrong? Could have done things better.

Uh, so there's that list of serious further offences. There were some things that were published. Government statistics.
And so we looked at those options and we talked to lots of people about internally about what, you know what, what, what does a serious offense mean to you?
We used my, you know, I've got a spreadsheet that's got like over 3000 criminal offences on it

and how do we categorize them for predictive purposes. And so we use that. And so eventually, you know, we brought all those things together through a working group and sort of came here this.
You know together and said, OK, this is consistent, these are the principles of what serious offenses are level like, that's satisfied. One of 10 criteria.
We need to check that it kind of works actuarially. So we ended up saying that terrorism and neglect child neglect offences, we agreed that they're serious, but we can't produce use an algorithm that will predict them. So it went out there saying RSR is a predictor of serious offending except for terrorism and neglect. You need to use your professional judgment for terrorism and neglects we're not going to pretend that the algorithm works for them when it doesn't.
Uh, yeah.
And so that's the kind of process that we went through. And then you have.
That there were things like certain occasion tool you got an RSR score.
If someone above what R score, does that make them someone who should be retained for the [redacted], and so then you'll you're doing creating models and so you're say, OK, who's a mapper offender who's a high risk of serious harm offender?

And who's got an RSR score above certain thresholds? If you run it this way, you'll get 30,000 people in the [redacted], or 50,000 people in a [redacted]. And so here are your various options. How big do you want the [redacted] to be? That kind of thing.

Now I can tell you we did all these options. It didn't quite work out that in practice like that in practice because having.... When decisions have consequences they didn't have before, sometimes it changes the decision. So people were probably doing their risk ratings. You know, after the break up, different tail, they did them before. So [redacted] ended up too big and the CRC is too small. And that's partly why they failed, because they're overheads were too high for the getting payment for the volume.
So, but yeah, that's the process. And so there are certainly a couple of minute meetings with [redacted] who at the time was the Under Secretary for [redacted]. So you know that you've got the ministers and the government department. You've got the Secretary of State at the top, and then you've got some ministers of state and then some. Yeah, I think he was Minister of state. And then you got some undersecretaries. So you got that three level structure. So grading set.

The strategy, but he I never met [redacted] and some people did, but he was at the top and then you had the lower level ministers and So people were would work up options, run them through the internal hierarchy. People like [redacted], and then it's like, yeah, OK, we can put this still a minister, you write a paper to the Minister, then you go to a meeting with the Minister while they ask you questions about it and tell you to go off and do something differently. And eventually the deadline is hitting and you have to agree on something. And that's the policy.
Basically.

0:25:10.850 --> 0:25:22.320
Ali Gordjahanbeiglou (Research Student)

That's that's great. Awesome. Thank you so much, , but I I think you also mentioned, I mean you covered the next two questions, how this participation of different stakeholders.

0:25:22.770 --> 0:25:23.180
Data Scientist 1
Umm.

0:25:23.200 --> 0:25:31.920
Ali Gordjahanbeiglou (Research Student)
Has being managed, so I'm gonna move to the people management issues HR issues, which is kind of like more. I focus my research aim.

0:25:31.220 --> 0:25:34.480
Data Scientist 1
Hmm. Ohh yes I did say there were several questions about that, yes.

0:25:34.580 --> 0:25:49.930
Ali Gordjahanbeiglou (Research Student)
Yeah. So up in your opinion, what's HR involved in making these decisions and making like considering the adoption of these predictive tools, HR, HR professionals in any way, whether involved or not really?

0:25:50.290 --> 0:26:1.130
Data Scientist 1
I don't think they were. No. So there were big HR consequences of what was going on with transforming rehabilitation. But those would have happened even if there weren't an actuarial tool.
You know, you had thousands [redacted] staff changing employer basically and having to determine what their terms and conditions are, were and how much flexibility the new employers would have, how much QP [redacted] people would have and things like that, I don't know to what extent they were sort of making up new rules and just enforcing new goopy stuff. I don't know the exactly what people got in terms of new contracts. So that was a consequence of the decision to Split the probation service. So even if they had just used exclusively Mechanistic rules that didn't involve a risk prediction algorithm, or if they'd and or clinical rules like the rosh.
Umm about how to split up the offenders then they would still have to do all of those things, so I can't remember us having engagement from HR people about how that was Done, no.

0:27:6.100 --> 0:27:9.910
Ali Gordjahanbeiglou (Research Student)
I see. OK. Thank you so much, but again.

0:27:10.680 --> 0:27:23.810
Ali Gordjahanbeiglou (Research Student)
How HR was contributing towards this change in the processes? Considering this, this algorithm this not algorithm, but I'm going to refer to this as a risk predictive tool but.

0:27:22.440 --> 0:27:25.360
Data Scientist 1
Yeah. Yeah, so.

0:27:24.570 --> 0:27:27.260
Ali Gordjahanbeiglou (Research Student)
How? Yeah, yeah. Yeah. Please. Please.

0:27:26.350 --> 0:27:34.140
Data Scientist 1
Yeah. So yeah, as I said, I just don't remember them being amongst those stakeholders cause it's really not an HR. It's a matter of clinical practice, not, not a matter of HR or how you know what you count as a serious reoffense and so forth. It's Uh, their implications for how people do their jobs. But yeah, it's more when you get into things that will affect You know will affect workload. So there's actually an interesting process called Tiering which I realize we weren't here to talk about, but.
How much time you spend on managing a case?
There's some structure to that so that managers can understand. The amount of workload that their staff have got cause a typical probation practitioner will have does manage dozens of people and that means that the lowest risk offenders actually you, maybe they get a 15 minute phone call once a month or something like that.
Uh, whereas the very highest risk people. Yeah, could have several hours of input per week. Uh, and so successively over the years there have been various structures that called tiering about putting offenders into these strata or now on a sort of four by four grid actually for the type of the amount of resource they're likely to need. And so the RSR is now part of the tiering algorithm. Again, I have that the direct contact there, but the people who were Running the tearing policy You know you have these implications for your setting up this structure. It has implications for workload because there are expectations about how much work you do with the people in each tier.
And then that gets into, yeah, workload management and notionally there's so many hours in the week and but in reality and probation it's it tends to be the probation officers have got You know their workload management percentage is over 100.
You know, once. Yeah. Is it 110 is 120, is it 150? You know, at what point does someone you know the rules step in and say somebody can't have That you know something over 150% at all, or they can only have it over 130% for a short time or whatever. I'm making those numbers up. I don't exactly know, but that's the kind of area where you have a mechanistic process. What are the tiering rules that can incorporate?

Some, you know, they incorporate RSR scores as an algorithm and ultimately it has an HR implication. I see. Yeah, I don't think it's. It was off the topic, yeah. But he was he was actually, yeah. Yeah, it's it's a. It's a bit of a shift of topic, but yeah, you have, you know you have something like when you have something like RSR gets used in the organization for various means which are. Not it's not inappropriate to use it in that way, but it's not how it was originally designed. What it was originally designed for, and you, yeah, you end up in conversations, OK, what's the appropriate way to do it?

0:30:45.290 --> 0:31:6.160
Ali Gordjahanbeiglou (Research Student)
That's great. Thank you. So I presume the trainings, the L&D requirements was kind of provided by you and your team for the those people who are responsible for monitoring to deal with this new technology. So they need they need that, I assume they need that some training, some preparations to interact with these.

0:31:10.230 --> 0:31:10.540
Ali Gordjahanbeiglou (Research Student)
OK.

0:31:7.580 --> 0:31:15.480
Data Scientist 1
They got basically guidance notes. So we had a, we had an implementation manager and. They sort of both project managed the digital support that there was for it because there was there's they have always been calculated. People have never had to do this using a pocket calculator or
Anything like that? The algorithms too complex for them to do that, so some additional support to serve and algorithm, and there was Also, these guidance notes are written, so we sort of wrote these guidance notes cooperatively and they decide, you know, they looked into the kind of information that people wanted and I'll test it out for you people with the scale of change that was occurring at the time of transforming rehabilitation, the split of services, that was what it's called, TR transforming rehabilitation. There just wasn't time to give people conventional training in every aspect of what was changing because so much was changing.

0:32:17.960 --> 0:32:20.610
Ali Gordjahanbeiglou (Research Student)
Thank you so much. Uh, that's. That's a very good answer.

0:32:26.930 --> 0:32:28.190
Data Scientist 1
Yeah, whatever it is. Yeah.

0:32:28.950 --> 0:32:30.60
Data Scientist 1
Yeah, yeah, that's OK.

0:32:21.10 --> 0:32:50.350
Ali Gordjahanbeiglou (Research Student)
Uh, uh, but uh, now focusing on the AI. Sorry, the predictive algorithm. I'm kind of used to saying AI, but yeah, apologies. Also focus on the ethical challenges, ethical issues of this particular technology in your opinion, where these ethical issues discussed, because technically there should be something. I mean, in theory there are some discussions in practice as well and we have we have this.
Topics in all over the news, but what these things were? You mean your opinion, whether discussed or?

0:32:56.940 --> 0:33:9.90
Data Scientist 1
Yes. Yeah, they, you know, would always be formally minuted or whatever. But, you know, there were processes that we would go through. So for example.
Umm, if you were working in [redacted] in the [redacted] then the last thing I heard their equivalent of our risk predictors would include somebody'd country of birth.
So it could be, yeah. Obviously the Netherlands or often you'd have like, uh, Surinam or the Netherlands Antilles or.
Um, Indonesia or you know all that you other European country or other country and that is a risk factor in their algorithm?

Now you know that effectively means that, um, that, that that ethnicity is in their algorithm. It's not a perfect proxy for ethnicity, but You know, uh, we've got nationality and we've also got ethnicity in the data we have. But I have always refused point blank to put those things in an algorithm because we are, you know, it's treating members of a group, as if that situation applies to them as an individual, which is an untenable assumption, of course, we have concerns about upstream bias in the criminal justice system, that what goes on in the police and the courts.
May affect people. We can't get rid of that entirely because people come to us with a criminal record, a formal criminal record, and we have to use that in prediction. You have to use that in prediction. It's a norm.
It's enormously predictive, and it's the best, uh, well, that plus some of these age are the best single predictors. So again, age is a characteristic that people aren't responsible for what age they are, but it's hugely predictive of serious non sexual violence in particular offenses or bounding and homicide are just massively correlated with age. The risk of proposed presented by an 18 year old offender is about twice the risk presented by 27 year old offender.
And then the right to decrease slowed. So you have to have those things in and algorithm. It does mean that it's like not fully individualized. The ideal would be that you would have an unbiased professional assessment of somebody's substance use and anger management and their, you know, employment and leisure time issues and their cognitive issues and those other things that the psychology of criminal conduct tells us there are associated with offending behaviour. And ideally, your algorithm would just include those factors that they can Reason that they held responsible for without getting too deep into the philosophy, free will and so forth. But you know.
I've tried that.
If you have an algorithm that just uses those factors, it isn't really isn't very good at predicting. It's not terrible, but it is very considerably worse than an algorithm that includes age and criminal history and gender. To a lesser extent, although gender matters quite a bit with serious violence and sexual offending, which are the things go into RSR.
If someone forced us to do our algorithm of all reoffending, including like shoplifting and stuff like that, take gender out, we could do it and the performance wouldn't for much.
But yeah, if you leave gender out of the serious reoffending algorithm, then you produce big overestimates for women because women Commit those offenses far less.
So You have those issues there that you have various things that are legally protected characteristics and we decide that the risks and the you know the bad consequences associated with including ethnicity or nationality are too great compared with the predictive benefits. And I for the public protection benefits you get. So we leave those out, but we decide that actually age and gender is really important to know about Those in prediction and so yes, it is treating an individualism as a member of an age group or as a representative of their gender. But we, you know, the They the discriminatory factor there is, is lesser you know is lesser and the benefits the public are greater. So essentially we have to draw a line and
You know, there's a degree of subjectivity there in the way those decisions are made, because you can't we can write down precisely how much the algorithm gets worse if you leave something out.
But then Exactly what happens in society as a result of the algorithm being worse? It's, you know, you have to make assumptions and what bad things happen into an individual as a result of being misclassified. You have to make assumptions.
So there's there is that aspect to it. So that I think is the biggest ethical factor that we Have to deal with there is some really interesting stuff that happens in risk prediction

It sounds technical, but.
And we don't have a perfect answer for it, but you may have heard about rape and how difficult it is to get a conviction for somebody who has, you know, has is a reported “rapist”.
So part of the algorithm, uh, the RSR algorithm is serious, non sexual violence. If someone commits an offense of wounding or homicide, more likely than not, they're gonna get convicted for it.
About 80% of homicides end in the conviction and for wounding I think it's less, but it's still pretty high, whereas you know 1 2% of rape offenses end in a conviction.
So we got serious non sexual violence in the algorithm and we got rape and your image is frozen. So can you still hear me? Yes, you're back. You're back.

0:39:2.550 --> 0:39:4.40
Ali Gordjahanbeiglou (Research Student)
Yeah, I was. There was a bit of him.

0:39:5.250 --> 0:39:5.880
Ali Gordjahanbeiglou (Research Student)
Yeah.

0:39:7.580 --> 0:39:9.580
Ali Gordjahanbeiglou (Research Student)
It was a bit of a poor network quality.

0:39:4.530 --> 0:39:20.410
Data Scientist 1
Yeah. Yeah. So we have RSR's. There's actually, there's. There's. Yeah. So there's actually 3 little algorithms that go into the the what we call the RSR algorithm, there's serious, non sexual violence, there's contact, sexual reoffending and there's indecent images of children.
So yeah, most serious, non sexual violence does lead to a conviction and so.
The base rate that you see in the RSO is quite close to the base rate in real life maybe, yeah, maybe you know, maybe 1 1/2 percent of our offenders go on to wound or kill somebody, whereas we, you know, in our stats it's 1%.
In our stats for men convicted of sexual reoffending, uh, the contact sexual rate is maybe 1% / 2 years. But that's the proven reoffending rate. Nobody knows what the true reoffending rate is. We've got these statistics published in the rate review where you can see the number of cases.

And how you know you can have 100,000 reports to the police and then you end up with 1500 men getting convicted of rape. And you can see the drop off at different stages of the process. But because we don't know. To take it to extremes, we don't know if there's 100,000 rapists committing one rape each, or if there's 1500 rapists committing 65 rapes each. Yeah, obviously, it's neither of those. It's somewhere in between. But where? In between. Yeah, I've read international academic papers where people are arguing about how to try to figure this out, and nobody has a satisfactory solution. So we don't know the extent to which the contact sexual part of the RSR algorithm is sort of an underestimate. So we're taking And again, indecent images of children. We've got the number of men who were convicted of it. So you can get a rate of indecent images conviction and we can do a little algorithm on that for our men convicted of sexual offenses.
But we don't know how many of them are actually perpetrating that offense again.
So you've got those three types of offending and.

On one side We're adding the predictions up, so we're treating the three types of reoffending as of equal seriousness.

Whereas in reality, you know, um, indecent images of children offending is horrible. But most people would say that homicide is more serious than that.
Yeah, go get more serious in homicide, right? So on the one hand, we're adding the free up, um. Which arguably deemphasizes the uh, the really, really serious offenses. You know, the worst rape offenses and the homicide offenses. But on the other hand, the serious non sexual violence bit is probably quite accurate. Quite well calibrated cause most of them get convicted, but the sexual offences. Ohh really under calibrated cause most of the perpetrators don't get convicted. And again we're just adding it up because we don't have a better answer. But I would say it's an ethical issue because in an ideal world we want to correct for that, and it would be better if we did. But yeah, then would we still have them all up or would we apply some sort of weighting?

You it's difficult to have these conversations.... even if we had the data to do that with, it would be difficult to have the conversations that would be very, very fraught.
As it is. We're doing something, we put it out there, we say what it is and you know people can then make up their mind what to do with that information. So to the extent that people actually read our briefing notes and things like that and know what the RSR does, you know, we're trying to tell them so to the extent that that's information is getting to them, we're being transparent, not with the exact form of the algorithm, but we are being transparent about what's going on.
And Yeah, that's kind of. It's fulfilling our responsibility, but it we know it's uncomfortable because we know we're in an imperfect situation.
So I think that's some of the major ethical stuff that uh plays on my mind.
Umm.
Yeah.

0:43:26.990 --> 0:43:52.910
Ali Gordjahanbeiglou (Research Student)
That's great. Thank you so much. I I believe the next question kind of overlaps with your last bit of your remarks. So considering OK the the predictive algorithm has an output for you, there is a there is a decision coming from this predictive algorithm. So how do you measure this performance or in other words, how do you consider this as a just decision? Is it a fair output? How do you measure this?

0:43:53.920 --> 0:44:2.750
Data Scientist 1
Yes. So there is an established process for this. My colleagues, um, probationary data science are actually gonna be Revisiting this Soon this, this second half of this year actually. I mean, I actually worked on it now, but it will probably be published like around the end of the year.
And so basically what they do is they get a cohort of people on the prison, the probation caseload in the community.
They establish what they're like in appointing time, so we're gonna take the caseload in the community on the 30th of June, 2018.
We've got data from the probation caseload system from the offender assessment system and from the police national computer that tells us what they were like until that point. So we've got a calculation of what their RSR score was at that point in time.
That police national computer data also actually we took the data um late last year so.
We can follow people forward for three years. Uh, if you like to mid 2021 from that 2018 point.

And so then we can see we can look at that record of offending.
What people are convicted of, and you don't just have the conviction date, you have the the what is understood to be the date of the offenses. So we're then looking in the data for offenses that were committed after the 30th of June 2018 and.
Within the next the two years, so 30 for June 2020 and then we've got data for about another year and sometimes it takes time to bring people to court and convict them so they could be convicted in the third year, so long as they committed it in the second year.
And so basically you then got a data set, so we had everybody's background information, including what there are ASR score was in 2018 and then we have the information. Did those seriously reoffend in the next two years?
We also have some other stuff like did they get sent to prison for burglary or something, which makes it impossible for them to then commit a serious events after that one more as impossible.
I'm the bad stuff. Happens in prison, obviously.
But so the you know the details stats is is more complicated than that, but essentially you've got everyone score and then you've got whether they seriously reoffended after that. And so you want the rate of serious reoffending to be far higher amongst the people with the higher RSR scores compared with the low ones. And there are a number of metrics that are being developed in risk prediction.
Uh, methodological literature, which is, you know, it's not just in our field. There's a huge, huge field of medical statistics which is far, far more extensive than the reoffending risk prediction literature.
So there is a lot of methodology that's being developed.
On the best ways of doing these measurements. We tend to use something called variously the concordance index or the area under curves AOC and the area under curve. If the predictor is useless, then the score is 0.5. If the predictor is perfect, then the score is 1.
It's like if you if you have this data set that we had, you're looking at after the fact, you got perfect information about what happens.
Put all the put all the people who did seriously reoffend him one box and put all the ones who didn't in another box pick out a record from each box. The error under curve is the probability that the person who did seriously reoffend had a higher score on the predictor than the person who didn't.
And so actually when they calculate it, they do all the possible combinations. So you've got 10 people in this box, 100 people in this box. And so you've got 1000 comparisons. And so the more the more comparisons that work.
The higher the AUC, the more comparisons that failed, the lower the AUC. If there are scores were just random, then half the comparisons would work and half of them would fail. So that's why 1/2 is the baseline.

And you never gonna get an AUC of 1 because that would require that actually all the reoffenders were bound to reoffend and all the non reoffenders were bound to not reoffend because otherwise you'd have some sort of noise, you know, some luck in reality people are like on a continuum. So some of the people with a score of nought point 1% will reoffend and Yeah, actually a higher pretty high RSR scores like 10%, so only one in ten of those is going to reoffend. So. So you know, there's a lot of noise in the system. So we're happy if we get something between .7 and .8 .
Uh, so yeah, if you get .75, then it's kind of halfway between You know raw, you know, complete

chaos and chance and perfect prediction. So that's how that that's roughly where we tend to end up in predicting serious reoffending. Predicting any reoffending is a bit easier. Predicting sexual reoffending is more difficult because you have so little information to go on.
Because they tend to have only been convicted or on one occasion of sexual offences, possibly of two, and there are reoffending behaviour, doesn't help very much. Um, so you've got, like, sexual reoffending rates in there and their age. So it's really difficult.
Yeah. And so that's basically how we do it.
Yes.

0:49:28.950 --> 0:49:55.300
Ali Gordjahanbeiglou (Research Student)
That's that's great. Thank you so much. So I think again, those measures, those protocols concordant, as you mentioned, you were referring to we, I mean, if I'm not mistaken, please correct me if I'm wrong, but I think you mentioned, yeah, you're referring to data science team. If there are other people involved in this, I mean, in forming this sort of or adding further insights to these protocols or to this concordance?

Other stakeholders in this or is it just purely related to the data science team?

0:50:1.880 --> 0:50:7.230
Data Scientist 1
Umm, so the data science team is at the heart of actually doing that measurement.
And so, you know, and there have been various people who've been involved over the years. It's just, yeah, you know, it's just on the on the constant. I'm the one who stayed in the team.
We work a lot with the public protection group of [redacted].
Umm, so we're having discussions with them about our program of work and sort of what should we look at next.
So there is the headline work of doing a big revalidation like what I've just describing him, getting a number that says how well it works, and also, you know, looking at that for particular subgroups of our vendors or particular offending outcomes of interest. But then there are also topics that we could explore around the details of the scoring rules.
Umm, so one that's coming up at the moment for example.
If you've got a sexual that men can man convicted of sexual offences and he gets done for not telling the police his new address or something like that, that's a criminal. That's a sort of a sexual offense because it's a breach of his sexual offending reporting requirements. But what do you do with that? in the algorithm? Do you ignore it, or do you count it as a sexual offense in his criminal history? Or do you sort of count it as half an offense if you like? So they want us to investigate those answers. So they've got an evidenced way.
And space way of telling the practitioners what to do to get the best.
Best score and therefore the best Baseline for their risk of serious harm decision.

Because ultimately the public protection group, you know, they believe in the risk of serious harm process, they know it's flawed. But they say it's vital to have that risk of serious harm, professional judgment process.

But they do want the actuarial baseline to it to be as good as possible.
That's all.

0:51:59.640 --> 0:52:6.120
Ali Gordjahanbeiglou (Research Student)
I see. Thank you so much. And apologies if there was a bit of background noise. I think some of the undergrads are very happy about the weather today.

0:52:6.580 --> 0:52:7.30
Data Scientist 1
Yeah.

0:52:7.110 --> 0:52:7.680
Ali Gordjahanbeiglou (Research Student)
Apologies.

0:52:9.90 --> 0:52:12.610
Ali Gordjahanbeiglou (Research Student)
Thanks so much. I mean there are just two more questions. So.

0:52:13.750 --> 0:52:39.420
Ali Gordjahanbeiglou (Research Student)
If I'm not mistaken, you're also collaborating with Alan Turing Institute regarding the implementation. Ethically ethical side of the AI as a external body. So if I'm not mistaken, I think [redacted] pointed that out. So to what extent this adoption and implementation of your technology, your predictive tool is in line with their protocols with their Viewpoints.

0:52:42.560 --> 0:52:43.10
Data Scientist 1
Umm.

0:52:44.520 --> 0:52:55.480
Data Scientist 1
I must admit that I haven't touched in touch base with that in a little while. I was quite involved with that pre-pandemic and responsibility for it's gone elsewhere.
I think it would very much be swinging back to me if we were introducing a new algorithm or if we were significantly changing our algorithm.
So yeah, but I know that, you know, they created these, um, these, the these criteria for the kind of characteristics that the algorithm should have. So yeah, I'm not gonna pretend to be right up with it right now, but I know we'd be going through that again if we make a big change.

0:53:23.90 --> 0:53:38.970
Ali Gordjahanbeiglou (Research Student)
OK. Very well. Thank you so much. Just one last question, So do you believe that the implementation of this risk predictive algorithm has been a success [redacted]? What's your reflection and what's your?
Uh. Other your team members? Reflection.

0:53:42.600 --> 0:53:57.180
Data Scientist 1
Yeah. I mean, I think my team members would defer to me on the basis that I'm the person who's been around since 2014. Umm. So I would say that it's a partial success, you know, on a on a technical level, it's been implemented.
You know, sort of mathematically correctly and all of that and the scores that are that

practitioners calculated are recorded and made available for analysis and so forth. And for you know, reporting back to them we have ,,.Yeah, we have tools that either straight their caseload to the managers and help them make resourcing decisions, things like that. So it's part of that ecosystem. It is also recognized by public protection group in that brush guidance and telling people to follow this process that starts with the RSR scored progresses on.
However, as I mentioned earlier, the communications aspect of it to the staff in the front line has been, you know, essentially a failure because we know from this user research as well as less formal feedback over the years that.
Uh. A high proportion of operational staff, you know, really don't know what the RSR score actually represents.
And also there is this mistrust of actuarial methods. They assume that actuarial methods will be far inferior to their professional judgments on the basis of things that are sort of common sense but wrong. Like I know the offenders so much better. So I can take all this in additional information into account. So my predictions were much better than what the actuarial says. People would think that I understand why they think that, but it's been proven Many times over, many fields of study, not just reoffending, that that's not the case, that algorithms are more consistent and that are So. Yeah. So there is that gap between the technical side of it and the sort of operational reality the RSR score is not as influential as it should be in how people actually make their decisions. It is a formal part of tiering, and it may well in the future be a formal part of criteria for which Umm offenders go on structured interventions to address their offending behaviour. Part of the targeting criteria for that. So in a sense, you know, the practitioners can't dodge that.
You know, and so that is in the, you know, there's is very well established principles that you should use risk of reoffending as part of the criteria for who goes on these programs. So if it gets into that, then that makes it more of a practical success because it will actually be influence it what people do.
Three, sorry. Some building going on out there.
Yeah, so you get the idea this this gap between what it should be doing and the impact of that has in the real world. And so at the moment, it's only part way there.
And you know the cause of that being this sort of gap between [redacted] understanding of.
First, the information that reaches them, and secondly there sort of back their background and the trust they have in these instruments and the, you know, the extent to which they know about the evidence or belief about the evidence. If they know it, you know, those are the factors.
You know, it's not their fault. It's the organisation. You know, we recruit, we don't recruit people, mathematicians to be [redacted]. We recruit People who are good at working with difficult people. So we have to give them more information in the right way and more support to help them make the right decisions.
That's think, what about say?

0:57:33.30 --> 0:57:39.780
Ali Gordjahanbeiglou (Research Student)
Yeah. Thanks so much. Thank you. I really appreciate it. If there are any final comments, final remarks, please feel free.

0:57:40.250 --> 0:57:41.510
Data Scientist 1
Yeah. No, I'm.

0:57:40.980 --> 0:57:44.620
Ali Gordjahanbeiglou (Research Student)
If you want me to stop the recording, if there's something again off the record, but.

0:57:44.60 --> 0:58:0.850
Data Scientist 1
No, no, I think it's. I think it's fine. So although the RSR is not, you know as we've discussed, it's not a, it's not full artificial intelligence. I think a lot of this still would apply with a full artificial intelligence system.
What the advantage the RSR score does have, and something that I am going to be trialing with a uh, a colleague, hopefully in [redacted] over this year. Is that potentially you can explain the scores, you know the rules are not in scoring, rules are not impenetrable. You can illustrate them.
And so we hope to do that. If you are a very.
Complex closed artificial intelligence system. Then you couldn't explain to the staff why the offender has got the score they've got and this kind of comes into the ethics. It was always important to us to have a score where we could explain to the staff what's going on. The fact that we've been unable to do that until now is frustrating, but we want to the staff member to understand why the offender, why the offender got this score and that will help inform their judgment.
Umm. And so that's an important thing to do, and people choosing between artificial intelligence systems have to consider.
Whether the whether the user of the you know of the tool is able to do that. And yeah, I mean I guess that would have HR implications as well as to what you're asking people to do in their job.

The degree of professional autonomy they feel, you know that the people feeling that RSR is a threat to that and other actuaries are threats to their professional autonomy that I guess that's nature thing as well. So yeah, there we go.

0:59:25.180 --> 0:59:37.550
Ali Gordjahanbeiglou (Research Student)
Yeah, that's again. Yeah, that's interesting. Very interesting. I believe that black box nature of any type of algorithm is, is hard. I mean, to explain it to the public, but I believe this is this is doing a remarkable job.

0:59:44.660 --> 0:59:44.990
Data Scientist 1
OK.

0:59:38.80 --> 0:59:47.220
Ali Gordjahanbeiglou (Research Student)
Uh, I think it's safer society for, for the people. I'm just gonna stop the recording and then we will. We will say farewell.