UNIVERSITY OF TARTU

Faculty of Political Sciences

Johan Skytte Institute of Political Studies

Emre Agirbas

Algorithm of Interest: A Qualitative Analysis of the Public Debates in Dutch, UK, and Australian Cases of Algorithmic Decision-Making Scandals

Master's Thesis

Supervisor: Vincentius Martinus Franciscus Homburg

Tartu 2024

Authorship Declaration

I have prepared this thesis independently. All the views of other authors, as well as data from literary sources and elsewhere, have been cited.

Word count of the thesis: 20069

Emre Yasin Agirbas, May 20, 2024

Non-exclusive licence to reproduce thesis and make thesis public

I, Emre Yasin Agirbas, (personal code:39509260167) herewith grant the University of Tartu a free permit (non-exclusive licence) to the work created by me [Algorithm of Interest: A Qualitative Analysis of the Public Debates in Dutch, UK, and Australian Cases of Algorithmic Decision-Making Scandals], supervisor [Vincentius Martinus Franciscus Homburg],

- reproduce, for the purpose of preservation, including for adding to the DSpace digitalarchives until the expiry of the term of copyright;
- to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until the expiry of the term of copyright;
- I am aware of the fact that the author retains the rights specified in p. 1;
- I certify that granting the non-exclusive licence does not infringe other persons'intellectual property rights or rights arising from the personal data protection legislation.

Abstract

This thesis examines scandals arising from algorithmic decision-making systems and the ensuing discourse, focusing on the concept of fairness in public debates. Three cases from Australia, the United Kingdom, and the Netherlands, where algorithmic systems were implemented in public institutions, are analyzed. Through qualitative research utilizing thematic analysis, textual data from e-petitions and media documents is thoroughly examined.

The study reveals procedural and distributive justice as central themes in discussions surrounding fairness. Despite extensive exploration of fairness in political philosophy, empirical investigations remain limited. This thesis seeks to address this gap by empirically investigating the role of fairness in algorithmic decision-making scandals, highlighting its significance in contemporary discourse.

Through the application of justice as fairness theory, this research contributes to a deeper understanding of the challenges and implications of algorithmic decision-making systems in public institutions. By examining real-world cases, it sheds light on the complexities of fairness in algorithmic governance, offering insights that can inform both policy and academic discourse.

Table of Contents

1. Introduction

1.1.Background

In 2011, more than 200 teachers in Washington D.C. lost their jobs due to their allegedly poor performance, despite some having received outstanding in-person evaluations (Colclough & Lappin, 2023). This action was part of Mayor Adrian Fenty's committed efforts to reform the education system. In line with this aim, he hired an education reformer Michelle Rhee who created the IMPACT system. This system aimed to detect the bottom %2 and get rid of them so that students would only be educated by well performing teachers. However, the algorithm did not consider the potential inflation of the students' grades caused by their teachers who do not want to be in the bottom, or socio-economic conditions that surely affect the grades of students.

Some of the teachers who were dismissed were highly esteemed by both students and parents. Some of their unexpected stories were even reported by The Washington Post. Sarah Wysocki, who had been teaching at MacFarland Middle School to fifth-grade students, was one of them. She was also considered one of the best teachers by her colleagues, students, and parents; the assistant principal stated in her May 2011 evaluation that "*It is a pleasure to visit a classroom in which the elements of sound teaching, motivated students and a positive learning environment are so effectively combined*" (Turque, 2012). One of the assessments also describes her as "*one of the best teachers I've ever come into contact with*" (O'Neil, 2016). However, after a while, she was fired because her students' scores in math and reading didn't increase as anticipated according to an algorithmic performance measurement system called IMPACT.

The interesting part is that Sarah Wysocki quickly secured a position at a school in a prosperous district in northern Virginia that does not fire people based on any score-related metrics. The former school, which has relatively bad conditions compared to the one in northern Virginia, lost its one of the best teachers. People, in some instances throughout the world, have become creatures whose value or acceptance within a society is solely determined based on their score. The score that algorithms create and use to decide the fate of the person of interest.

While using algorithmic decision-making systems to reach the best performance sounds charming, in fact, it may possess some adversarial impacts in the broader sense. Since algorithms rely on data, it tends to keep the existing social structure. This makes it difficult to ensure that fairness and equality of opportunity prevail; moreover, it raises questions with regard to justice. The anecdote

highlighted in the introduction serves as a compelling example of the direct impact of algorithmic decision-making systems on public services. It is evident that the integration of algorithms in public institutions will extend beyond conventional boundaries, influencing service performance, quality, and adherence to public values. In this manner, the role of algorithmic decision-making systems in public institutions ought to be studied as empirical studies over philosophical methodologies in political science context.

The case study mentioned above provides a strong example that algorithms used to measure performance in education can lead to incorrect results. Teachers' careers should not be determined solely by the scores produced by algorithms. The effects of such systems on the social structure and understanding of justice should be seriously considered.

The case study, which gives such a message, actually reminds us that we need to question the increasing effectiveness of algorithms in our lives.

The role of algorithms in public services can affect not only performance but also service quality, performance, and compliance with public values. Therefore, the role of algorithmic decision-making systems in public institutions should be examined through empirical studies in the context of political science.

The organization of the paper is outlined as follows. The introduction provides an overview of the research topic and sets the context for the study. It introduces key concepts and outlines the significance of the research problem. Following the introduction, the research design and methodology section details the approach taken to investigate the research questions, including the research design, data collection methods, and data analysis techniques employed. Additionally, the section discusses any limitations and ethical considerations. Subsequently, the theory section delves into John Rawls' "A Theory of Justice," providing a comprehensive overview of its key principles and how they relate to the research topic. The literature review synthesizes existing research and theoretical perspectives relevant to the study, identifying gaps and establishing the theoretical framework for the research. Lastly, the research findings section presents the empirical results of the study, analyzing the data collected and discussing their implications in relation to the research questions and theoretical framework established earlier.

1.2.Motivation

The rapidity of advancements in the use of algorithms is related to the evolution of artificial intelligence (AI). AI has not only increased the role of algorithmic decision-making systems in our lives, but it has also become the most important driving force behind changes in our socio-economic life (Jiang et al., 2022).

In this regard, it has an enormous potential to alter the many aspects of our individual lives and, more broadly, fundamental building blocks of society. It also affects the means of production, and, similarly, class struggle also earns a novel identity. As Coeckelbergh (2022) argues in his book, The Political Philosophy of AI, regardless of its sub-field, technology itself cannot be politically neutral. This applies to algorithms.

That being said, the focus of my work will be on algorithmic governance in the delivery of public services rather than general uses. While algorithmic steering of policies and governance processes suggests that AI can influence politics generally, my interest lies specifically in the use of algorithmic decision-making systems. At this point, emphasizing that algorithms are not neutral will only be valid when examining the political effects of their use in the provision of public services.

Even if the algorithms were to be used in favor of disadvantaged strata within a society, it would still be "*symptomatic treatments"* (Coeckelbergh, 2022, p. 72) instead of addressing the underlying causes. When data is involved then bias, political leanings, or world views that people carry are injected into the related tech-product. It is mainly because of the fact that data can never be raw since it is collected, processed, and interpreted. This whole operation is not exempt from the inherent bias. Similarly, Dyer-Witheford, Kjøsen, and Steinhoff (2019) contend that AI could be regarded as the pinnacle of the alienation of workers within a capitalist context. Fuchs (2014) also claims that people have become data producers for big tech companies and people are not only exploited to get their data but also, they are targeted to sell the product. This unsettling relationship dehumanizes human beings and makes them a commodity in the market-based capitalist society.

Rafanelli (2022) regards AI as, rather than stripping power from humans, instead, cultivating human power with considerations of justice being pivotal to its responsible implementation. Even though it is debated in the literature that AI cannot be neutral, the injustice that is created in this manner is not inherent in AI itself, rather, it is a reflection of those who wrote the AI system's code,

the data they used, and those who decided to use them. In this manner, it falls upon us as consumers, programmers, and researchers to ensure that fairness is minded. Erman and Furendal (2022) stress that it is needed to acknowledge the increasing significance of examining the societal and ethical repercussions of developments in AI. Moreover, they state that an interdisciplinary approach is essential, and the last years have been witnessing increasing attention from philosophers with a penchant for technology, as well as engineers with a philosophical inclination.

While the concept has been deliberated from a philosophical perspective, there is still a necessity to enhance the literature regarding the empirical and analytical examination of fairness. In this manner, Schiff et al. (2022) examine failure in government adoption of AI with using a survey experiment to gauge how citizens respond to hypothetical situations where public values are not achieved. In their empirical study they examine the citizens' reactions when their government fails to ensure public values which are composed of fairness, transparency, and human responsiveness when AI is adopted. They argue that any public value failures in this manner cause consistent negative effects on government evaluations. Moreover, such evaluations go beyond the policy context and political ideologies. In another empirical study, Mead and Neves (2023) also examine public responses to AI implementation in the context of entrusting decision-making of which the cases are Robodebt and A-Level Grade Scandal that are also cases of this study. They draw attention to the concerns that non-human entities have replaced judicious decision-making by humans.

The pace of algorithms' advancement and integration into public institutions is unparalleled. As this evolution unfolds, academic literature adapts to reflect these advancements. However, the question of fairness concerning algorithmic governance remains a notable gap in empirical practices. As it is aforementioned, empirical/analytical studies are relatively scarce in literature compared to philosophical or theoretical approaches in this topic. This research endeavors to bridge this gap by conducting empirical analysis on the intersection of algorithmic governance and fairness.

1.3.Research Question

The thesis delves into the utilization of algorithms within public institutions for decision-making processes. Having said that, this thesis centers on the public discussions sparked by these applications, instead of analyzing the overall impacts of decision-making processes based on

algorithms. Public debates can clearly reveal the real-world consequences of algorithmic decision-making processes and the perception of these processes on society.

When addressing issues within the public sphere, it is imperative to prioritize the overall welfare of society, ensuring inclusivity and equity without discrimination against any segment of the population. Although this seems like a simple proposition in theory, real-life practices show that this may not be the case. This may arise both from the embedded prejudices carried by the data feeding the system and from the tendencies of the technocrats/politicians who manage the system. Therefore, it is important to determine whether these implemented systems produce fair results. Still, what is a fair result is a matter that has its own nuances and involves too many details to decide just by looking at the output. Thus, studying such debates can be useful in understanding algorithmic decision making. It provides an important window into understanding their impacts.

From this point of view, the research question that this study aims to answer is "*How is the concept of fairness addressed in public debates about algorithmic decision-making scandals?"*. This study seeks to enhance the existing literature through an empirical investigation into the relationship between algorithmic decision-making and fairness. While previous research predominantly addresses the connection between fairness and artificial intelligence (AI) from a philosophical standpoint, this thesis takes a novel approach by employing principles from political science and bolstering its arguments with empirical evidence. By adopting a qualitative methodology, specifically a small-n design, this research aims to delve deeper into the intricacies of the issue and offer diverse perspectives. Through this endeavor, the study aspires to contribute to a more comprehensive understanding of the subject matter, thereby filling a gap in the current literature.

1.4.Theoretical Orientation

In this study, Rawls's views on justice constitute the theoretical perspective, serving as a cornerstone for understanding the intricate interplay between societies and the concept of fairness. By delving into Rawls's theories, we not only explore the nuanced dynamics of societal fairness but also illuminate the diverse contexts in which these principles can be applied, thus enriching our comprehension of justice within broader socio-political frameworks.

Rawls's theory of justice has been extremely influential in the field of philosophy and has been widely discussed. Although it is an abstract theory, and primarily utilized as a philosophical perspective rather than empirical studies, Rawls' principles do not have to be limited to theoretical

discussions. Rawls' theory of justice as fairness holds significant potential for practical applications. Especially on issues of algorithmic bias and justice, Rawls's understanding of justice has provided an important guide and conceptual framework. Thus, it serves as a promising foundation for research on fairness in algorithmic decision-making.

Defining fairness in the context of algorithmic decision-making systems is a relatively difficult issue. There is a technical side to the issue and many fairness definitions that have been developed are context dependent. Due to the technical features of the designed algorithms, it may not be fair to a certain segment, or fairness may be achieved in a certain context. Rawls' theory offers a comprehensive analysis window at this point, and this window it offers is closely related to algorithmic justice (Barsotti & Koçer, 2022).

Rawls' conceptualization of the concept of justice as fairness and the principles he presents will serve as a fundamental perspective in interpreting the data. Moreover, delving into the intricacies of Rawlsian theory in Chapter 3 will not only provide a comprehensive understanding of his framework but also shed light on its implications for our analysis of the empirical findings.

1.5.Relevance

Algorithmic decision-making mechanisms are becoming increasingly common. This mechanism, which has the potential to affect our lives so deeply, needs to be examined both academically and socially. Public debates help shape the narrative and how narrative is shaped is a matter of power relations. The process of developing a discourse that includes citizens is also an assurance that citizens have the right to have a say in discussing technological developments. Only in this way can the public be involved in the course of technological developments. It is important to examine the public debate rather than the issue itself. Otherwise, decisions about us might be determined by a technocratic and rich segment of society who use a technical language we do not understand. Public debates hold particular significance as they serve to raise awareness among the people regarding ongoing issues, while concurrently shaping the discourse surrounding these topics from the collective perspective of society.

Involving the public in the process of AI-related developments or listening to their voices on such an issue prevents society from falling into the trap of techno-optimism. Techno-optimism is, in essence, the worldview that a better world can be achieved by combining technology with human passion and ingenuity (Danaher, 2022).

According to Acemoglu and Johnson (2023), the world has been infused with the techno-optimism of tech giants. They also draw attention to how dominant classes may impose their profit-seeking approach as technological development. Such developments are likely to benefit the wealthy minority who can control the means of production. To ensure that technological developments positively impact not only a small, privileged group but also the entire society, it is crucial to make the public aware of ongoing advancements. The way that ordinary people perceive or embrace the issue will likely impact the course of human and non-human intelligence relations.

In conclusion, the examination of public debates surrounding algorithmic decision-making mechanisms is paramount in shaping the trajectory of digital governance practices. By involving the public and considering diverse perspectives, we can mitigate the risks and ensure that technological developments align with societal values and interests. Ultimately, the way in which ordinary people perceive and engage with these issues will play a crucial role in determining the future of human and non-human intelligence relations.

2. Research Design and Methodology

2.1. Research Design

This study was conducted as a small-n qualitative study as it aims to provide comprehensive descriptions and rich analysis by examining three examples of public debates. According to Zimmerman (2022), small-n methods are characterized by an emphasis on detail. A researcher has to be able to see the environment that they are studying. The purpose of this method is to gain an in-depth knowledge of particular cases. Small-n methods prioritize attention to detail, and the essence of it lies in acquiring comprehensive insights into specific cases. Three cases were analyzed in the study, and two cases are combined based on existing academic studies and secondary data with a more in-depth study of petitions in relation to algorithmic governance. The SyRI case from the Netherlands and A-level exams case from the UK are going to be conducted as exploratory case studies. A more in-depth qualitative analysis will also be conducted for Australian citizens' petitions. It is aimed to conceptualize categories and attributes of fairness in real-life governance practices with our exploratory cases.

For the research question to be answered, this study utilizes thematic analysis method in its assessments of the three cases. It encompasses the analysis of textual data, including interviews, newspapers, brochures, or websites. Thematic analysis offers flexibility both in coding processes and creating themes out of these codes. According to Rigger and Sigurvinsdottir (2017, p.33), "*Thematic analysis is a method for analyzing qualitative data that involves searching for recurring ideas (referred to as themes) in a data set".* In thematic analysis method, instead of quantification of the textual data it is aimed to reveal the existing pattern through created themes.

Thematic Analysis is widely regarded as the most suitable method for studies aiming to uncover insights through interpretation since it offers a structured approach to analyzing data (Alhojailan, 2012). Thematic analysis is flexible in a way that a researcher follows a relatively consistent path from deciding on research questions, choosing resources, and creating a coding frame to the presentation of the findings. This flexibility allows researchers to recognize the underlying meaning. Namey et al. (2008, p.138) states that

> *Thematic analysis moves beyond counting explicit words or phrases and focuses on identifying and describing both implicit and explicit ideas. Codes developed for ideas or themes are then applied or linked to raw data as summary markers for later analysis, which*

may include comparing the relative frequencies of themes or topics within a data set, looking for code co-occurrence, or graphically displaying code relationships.

In this method, a researcher may initially disregard certain details in the presented document, focusing instead on identifying and extracting data pertinent to the research question. However, as the analysis progresses, there may arise instances where previously overlooked details become relevant, prompting the researcher to revisit and incorporate them into the analysis framework. This iterative approach allows for a more nuanced understanding of the data and enhances the depth of the research findings.

2.2.Case Selection and Data Collection

In an effort to find the most suitable cases for research question many similar cases are taken into consideration throughout the world. Given both the research question and the methodology of the study it was crucial to cover the topic from different angles. Cases should reflect the insight and outsight debates surrounding the incidents. With this aim, perfect combinations to meet these criteria I selected the Robodebt case due to its rich dataset, offering substantial insights into the matter. Similarly, the other exploratory cases were chosen for their ability to ignite public discourse and debate and they completely occupied the agenda in their own countries and, in a sense, have become the subjects of discussions that transcend borders. Some of the possible cases either did not provide the necessary data or the cases were far from creating a public discussion environment.

Cases are not selected based solely on the country profile; instead, my approach involves researching numerous similar cases. I take into account whether a particular case has sparked public debate and attracted significant societal attention. Additionally, I consider the cultural, and socio-political contexts surrounding each case to ensure a comprehensive analysis that goes beyond mere surface similarities. By adopting this approach, I aim to capture the nuances and complexities inherent in each case study, thereby enriching our understanding of the broader issues at hand.

In the Netherlands, such an incident led the government to resign (Heikkilä, 2022) and in the UK it caused street protests (Shead, 2020). However, for some similar incidents, it was obvious that such research would be difficult due to the scarcity of textual data. All the cases in this study caused serious debates in the public sphere and triggered legal actions.

The common thread among these cases lies in their pivotal role in sparking public debates within the realm of algorithms and governance, ultimately shaping the trajectory of algorithmic governance structures within their respective countries. These cases have not only catalyzed widespread discussions but have also laid bare the intricate interplay between technology and governance practices. The nature of these debates, intricately linked to algorithmic governance, underscores their significance in understanding the evolving landscape of governance in the digital age. Thus, the selection of these cases was driven by their ability to provide rich and multifaceted insights into the complexities of algorithmic governance. Prior to their selection, criteria such as the richness of available data and the magnitude of public discourse were carefully considered, and all three cases utilized in this study successfully met these criteria, ensuring a robust foundation for our analysis.

On the other hand, it is essential to note that the selected cases are compatible with the chosen theory. One of the criticisms directed at Rawls is that his theory comes from someone who experienced American democracy, which Rawls, in a sense, "embraced" such criticism because his works are addressed to constitutional democracies with market economies (Stein & Harper, 2005). These countries, in general terms, are countries that have adopted a market economy where the rule of law prevails.

The cases that are analyzed in this study are the Robodebt case from Australia, the Systeem Risico Indicatie (SyRI) from the Netherlands, and the A Level grading scandal from the United Kingdom. These cases are about algorithmic decision-making systems that were used in public institutions. SyRI is a digital welfare fraud detection system that was later decided to be unlawful by the court due to its failure to adhere to the privacy rights outlined in the European Convention of Human Rights (Van Bekkum & Borgesius, 2021). The A-Level grading scandal occurred in the United Kingdom in 2020. Due to the COVID-19 pandemic, traditional A-level exams have been canceled and replaced by a teacher-assessed grading system. However, a controversy later emerged over the algorithm used by the exam regulator, which is called The Office for Qualifications and Examinations Regulation (Ofqual) to standardize grades. The algorithm led to significantly lower grades for some students, especially students at historically lower-achieving schools. This controversy caused widespread criticism, protests, and legal challenges. Eventually, the UK

government decided to abandon the algorithm-based grading system and allow students to receive teacher-assessed grades or administered grades (Bedingfield, 2020).

The Robodebt scandal is an Australian incident and specifically concerns the use of algorithm-based decision-making systems in public service delivery. This scandal emerged in 2016 in connection with the Australian government's use of an automation system in the distribution of social benefits, called Robodebt. This practice miscalculated people's income and created incorrect debts. People facing false debts created a massive outcry in society, causing many people to fall into financial and emotional difficulties. The Robodebt scandal subsequently sparked criticism and investigations for faulty debt collections and the algorithm making faulty decisions based on personal data. Many people were victimized, and the scandal caused a wide public debate (Rinta-Kahila et al., 2023). The Robodebt scandal was also investigated by a Royal Commission established by the Australian government to understand the effects of algorithm-based decision-making systems on people and examine the measures that should be taken to correct them.

The data collection process differs between Australian case and the other two cases. After reviewing the academic literature on the A-level exams and the SyRI scandals, this was further elaborated with a content analysis of media reports on both cases. The data for the case from Australia was gathered from submissions (e-petition) that were made to the Royal Commission, and Royal Commission's report based on these submissions. In Australia, royal commissions are the most authoritative method of investigating matters of public concern. The Royal Commission has broad powers to collect information necessary for its investigation. This includes the ability to call witnesses to testify before him and the ability to request individuals or organizations to submit documents as evidence. The Royal Commission has received a total number of 1099 submissions from 21 September 2022 through to 2 June 2023 of which 602 submissions are public ("Submission to the Royal Commission into Robodebt," n.d.). The Royal Commission has published all the submissions that were both relevant to the Commission's Terms of Reference and appropriate for public disclosure. Thus, those who are seen on the Royal Commission's website are examined and eliminated data according to their appropriateness. Within these submissions, individuals detail the impact of the Robodebt program on their lives. Additionally, they articulate the extent of their suffering.

2.3.Method of Analysis

In this study, a synthesis of inductive and deductive methodologies was implemented, reflecting a deliberate effort to combine the strengths of both approaches. By integrating these methodologies, we aimed to ensure a holistic and rigorous exploration of the research questions at hand. The inductive aspect of our methodology allowed for the emergence of insights from the data itself, enabling us to uncover novel patterns and themes. Concurrently, the deductive aspect provided a structured framework rooted in existing literature and theoretical frameworks, guiding our analysis and interpretation. This dual approach underscores our commitment to comprehensive engagement with the breadth and depth of existing scholarship, while also fostering innovation and critical inquiry.

As a foundational step, an initial coding scheme was developed to guide the subsequent coding process. Nonetheless, responsiveness to emergent themes from the empirical context led to the augmentation of the initial codebook with new categories. Consequently, both the coding process and data analysis unfolded iteratively, reflecting an ongoing refinement and expansion of analytical frameworks in response to empirical insights. The coding process was both concept-driven and data-driven.

First of all, an initial coding scheme was first created based on the literature and theoretical framework. These initial codes provided guidance in the coding process and in subsequently determining themes while analyzing the data from the cases, allowing for systematic examination of the data. During the coding process, emergent themes in the data, which were independent of the initial coding scheme, were also incorporated into the analysis and the eventual coding frame is created, which can be found at the end of this section. These initial codes are non-hierarchical, meaning there is no ranking or order among them. The initial coding framework was developed comprehensively for two primary reasons: to incorporate a wide range of possibilities and to accommodate the emergence of new codes during the inductive process.

On the other hand, to accommodate the inductive strategy, the coding approach developed by Strauss and Corbin (1998) was adopted in order to examine and categorize the data in detail. This approach envisages coding as a process that takes place in three stages: open, axial, and selective coding. However, in this study, the open and axial coding processes were carried out together, that

is, these stages were carried out simultaneously rather than sequentially. In this way, examining and analyzing the data was carried out in an iterative process.

In the open coding process, texts are read repeatedly and broken down into smaller pieces called units. These units of meaning are labeled with appropriate codes to capture the essence of the data. Codes are carefully selected to reflect the general structure and important themes of the dataset. In this regard, the coding frame is arranged as flat to grasp the meaning of the textual data better.

The relationships, connections, and structures between different open codes were revealed at the axial coding stage. In particular, the focus was on how different elements in the data set relate and interact with each other. Axial coding was used to transform the open codes obtained from the open coding process into broader categories.

The selective coding phase involves identifying the main categories obtained from the axial coding process. At this stage, the main themes in the data set are identified and focused on. Around the main categories, other elements in the data set are organized and given meaning, and results are organized based on major categories.

Initial coding scheme is presented in the beginning of research findings sections with the explanation of underlying logic.

The Process of Coding

- The initial coding scheme is created both based on Rawls's theory and literature review.
- Review of Data: In order to understand the general structure of the data set, it will first be reviewed, and important topics in relation to the research question will be identified.
- Open Coding: The process of separating the data into meaning units and assigning appropriate codes will be carried out. This process also involves re-reading of data.
- Categorizing the Codes: The obtained codes will be grouped according to their similarities and relationships, and a flat structure will be created.
- Development of the Coding Framework: A coding framework will be created that defines the relationships between categories and subcategories.

3. Theory: A Theory of Justice

3.1.Introduction to the Theory

In 1971, John Rawls published 'A Theory of Justice,' a seminal work that profoundly influenced political philosophy and continues to shape debates on justice and fairness. Rawls, to some extent, takes a position from a social contract perspective and builds his theory of justice in relation to this approach (Torbert, 1974). Social contract theory implies that there is an underlying agreement between the constituents of society, let it be individuals or institutions, to establish the governance structure. This approach has been voiced by prominent thinkers in political philosophy literature. Social contract theory provides a framework for how the basic structure of society and the legitimacy of government are achieved. This theory suggests an agreement between individuals to protect their natural rights and live together. Despite the varied methodologies of social contract theorists, they share several fundamental similarities. These similarities revolve around an effort to define the principles of legitimate political authority using a rational and materialistic framework. Moreover, social contract theorists ground their theories in specific descriptions of human nature (Seabright et al., 2021).

However, Rawls's approach to social contract tradition relatively deviates from others since Rawls put importance on fairness and rationality. Rawls's approach, unlike other social contract theorists, attaches importance to being fair and rational. Unlike other theorists, Rawls built his social contract theory on fairness and rationality rather than the interests of individuals or the power structure between them. In his view, social contract theory should aim to create a just structure to ensure the welfare of the weakest strata of the society. This is a unique approach that focuses on achieving social justice rather than the individual interests emphasized by other theorists. The common understanding of social contract revolves around the self interest of people and the preservation of their natural rights.

In its essence, according to Rawls, at the core of justice lies the basic structure of society. Basic structure, as he defines, "*is the way in which the main political and social institutions of society fit together into one system of social cooperation, and the way they assign basic rights and duties and regulate the division of advantages that arises from social cooperation over time*" (Rawls, 2001, p.10). For Rawls, the basic structure of society is central to his theory of justice. The basic structure describes how the major political and social institutions of a society come together as a system and

how these institutions assign basic rights and duties and regulate the distribution of advantages arising from social cooperation over time. The basic structure includes an independent constitutional system, legally secured private property, economic conditions, and even an institution such as the family. These institutions form the social structure in which individuals, groups or institutions operate in the background. In this context, the efforts of the individual are not as decisive as these institutions. The logic of the basic structure lies here. Moreover, justice deals with how fundamental rights and responsibilities are allocated within society and how benefits that are created from social collaboration are distributed by social institutions. These institutions can be characterized as the fundamental elements of the social framework, encompassing both the political constitution and the primary structures governing economic and social relations. The economic and social arrangements constitute a broad perspective, which includes cultural norms, private market or education systems, as well as legal frameworks in that society. The importance of these institutions derives from determines people's rights and their well-being within the society. How these institutions are shaped is highly related to the conditions of justice.

Before processing, two concepts of Rawls need to be laid out to completely comprehend the further elaboration. These are original position, and veil of ignorance concepts.

Rawls proposes a concept of original position as an imaginary device which could create theoretical scenario in which participants are placed behind a veil of ignorance. These rational actors are aiming to choose principles of justice for society. However, since they are behind the veil of ignorance, they lack the information related to their position within society. Without knowing their race, wealth, talents, or any other social characteristics it is ensured that no one is favored over another. The goal of the veil of ignorance is to ensure that no one occupies an advantageous position, thereby aiming for an expected fair outcome. Since the participants are rational actors then it is assumed that they will try to reach the fairest outcome while everything is both equal and unknown for them.

The aim of the original position is to determine base principles for justice that rational individuals from their unbiased position would opt for. These principles are expected to be suitable for everyone, regardless of their position in society. Thus, these principles of justice originating from the original positions focus on equality and the wellbeing of least-advanced strata of the society.

For Rawls, a rational actor would choose justice as fairness instead of any utilitarian approach in the original position (Lovett, 2011). Rawls's theory is often characterized as egalitarian because it restricts inequalities in favor of those who are least advantaged (Moon, 2015). However, simply limiting the inequalities experienced by the least advantaged may not be enough to understand how Rawls is positioned. Rawls's concept of justice as fairness is also a new interpretation against the utilitarian approach. According to Torbert (1974, p.462) "*He repeatedly contrasts his approach to that of classical utilitarians, such as Bentham and Mill. Rather, Rawls follows social contract theorists, like Rousseau, Locke and Kant, in emphasizing the natural rights of individuals*". He further states that his concept of original position is similar to social contract approach of theory's predecessors. Moreover, social contract is very useful in formulation of just principles, considering that justice might be very difficult for cenceptualization since every culture has its own understanding of the world. At this point, Rawls offers a solution to this conceptualization problem by introducing general principles of social justice (Orji, 2023).

Especially, when the original position is introduced, thanks to the veil of ignorance, Rawls argues that two principles emerge. These are the principles of equal liberty and principle of equality.

The main function of these two principles is to lead the way for the aforementioned social institutions, which helps the proper functioning of the basic structure.

3.2.Principle of Equal Liberty

Rawls articulates that individuals have rights inherently by virtue of their existence as human beings. Furthermore, these rights are inviolable and cannot be revoked under any circumstances. Compared to Rawls' second principle of justice, this first principle is foundational for establishing conditions of fairness. In his entire theory, this is where Rawls comes closest to other social contract theorists, given that social contracts deal with the initial conditions of society. Here, Rawls discusses factors that are inherent from birth, which, in a sense, are aspects that are naturally available to us.

Even though this state of being in nature is imaginative, people's access to them cannot be prevented and is their birthright, as long as their oppression does not restrict their access. These rights emphasize that each person is equal in possessing and enjoying fundamental freedoms. These fundamental freedoms are freedom of religion, expression, thought, and conscience. On the other hand, just because a person has these rights does not mean that they have the authority to restrict the rights of others.

The limitation of these rights can only be warranted when aimed at safeguarding these very rights in a broader context. Hence, these rights might be revoked from groups that oppress other groups.

3.3.Principle of Fair Equality of Opportunity

The principle of equal liberty bifurcates in two ways. While the first emphasizes the least advantaged strata of society needs to be prioritized, the second one focuses a bit more on the economic inequalities and equality of opportunities.

3.3.1. The Difference Principle

The difference principle, as formulated by Rawls, asserts that social and economic inequalities should be arranged so that they are both to the greatest benefit of the least advantaged members of society. It emphasizes the importance of ensuring that any inequality in society results in the greatest possible advantage for those who are least well-off. This is not to say that their life should be open to direct involvement. Rather, the system itself needs to be organized accordingly. Basically, if an individual or group can benefit from something it should encompass the lower class also.

On the other hand, such an argument might be sound implying the justification of inequalities. Inequalities can only be justified to ameliorate the conditions of the lower classes. For instance, in terms of distributive justice, if a local municipality aims to deliberate social welfare aids to university students, then it should keep in mind that not every student is coming from similar background even if they study in the same university. Any inequity in the distribution of these aids is not deemed unjust if it prioritizes those in the greatest need. The underlying logic here is that arbitrarily personal factors do not directly belong to us and should not determine our opportunities in life, which aligns with the concept of the original position. Factors that are beyond our own control such as talent, the family we are born into should not be a leverage in our position compared to other individuals.

Another aspect of the difference principle involves its consideration of future generations. It is sensible to regard the next generation as among the least advantaged strata of society, as akin to lower classes, they lack the means to defend their rights or improve their circumstances presently. Therefore, prioritizing them is essential, and it is incumbent upon all to remember their responsibilities towards them.

3.3.2. The Equal Opportunity Principle

The equal opportunity principle aims to provide an equal starting point for every individual in a society when it comes to education, healthcare services, job opportunities etc. This is where and how people can contribute to society and fulfil their potential. The equal opportunity principle is vital to social justice and an egalitarian society.

Rawls' theory proposes that these two principles be considered as the basic principles for organizing social and economic structures. Rawls' principles of justice profoundly affect the functioning of social institutions within society. For example, Rawls's principle of fair equality of opportunity imposes an obligation to protect and extend the fundamental freedoms of individuals. This principle provides a fundamental guide in guiding the policies and practices of legislatures, judicial bodies and other institutions. Likewise, Rawls's 'difference principle' promotes a just correction of social and economic inequalities in society. These principal shapes state intervention in areas such as tax policies, social services and education. These examples help us better understand the impact of Rawls's principles of justice on social institutions.

Based on these principles, we can evaluate whether the inequalities that exist in society are legitimate and provide a framework for what a just society should look like.

Regarding the relationships among these principles, the "principle of equal liberty" takes precedence: social institutions should not promote equal opportunity or permit economic inequality in ways that infringe on individuals' fundamental rights and freedoms. Likewise, the "equal opportunity principle" supersedes the difference principle: under no circumstances equality of opportunity can be compromised to enhance the income and wealth of the least advantaged (Follesdal, 2014). In a sense, this ranking among the principles actually summarizes Rawls' stance against the utilitarian approach in general.

3.4.Synthesis of the Equal Liberty and Difference Principles

In the previous sections, Rawls' theory of justice was examined, emphasizing the principles of Rawls. To summarize, the principle of equal liberty underscores the inherent and inviolable rights individuals possess by virtue of their humanity. These rights include freedom of religion, expression, thought and conscience, provided that they do not restrict the rights of others. However, these rights can only be restricted or revoked from oppressive groups to safeguard them in a broader context.

The difference principle relates to the view that social and economic inequalities should be arranged to the greatest benefit of society's most disadvantaged members. Inequalities can only be justified if they improve the conditions of the lower classes, so the system must be organized to prioritize those most in need.

4. Literature Review of Fairness in Algorithmic Decision-Making

The concept of justice, which was discussed from the perspective of Rawls in the previous section, will be examined in this section in the literature in the context of algorithmic decision-making.

4.1.Methodology of the Literature Review

This literature review aims to address AI implementation in public institutions. The main inquiry that this literature review seeks to address is how AI implementation in public institutions is discussed in literature and how or if, in relation to the research question, fairness is conceptualized. Although the research question focuses on algorithmic decision-making systems, the literature review intentionally utilizes the term artificial intelligence to encompass a wide array of studies. In certain instances, artificial intelligence and algorithmic decision-making are used interchangeably.

To achieve this objective, this study employs the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) chart. The PRISMA flow diagram offers a visual representation of the screening process, documenting the initial number of identified articles and enhancing transparency in the selection process. It delineates the decisions made at various stages of the systematic review.

The overall goal of this systematic literature review is to provide a comprehensive and reliable overview of current knowledge, particularly focusing on the conceptualization of 'fairness' in the context of public service delivery. By synthesizing existing literature, this review also aims to clarify and define 'fairness' within public service delivery, while also shedding light on promising research directions regarding AI implementation in public institutions. Through this process, the review seeks to contribute to a better understanding of 'fairness' in the context of AI-driven public service delivery and informing a coding scheme to be used in Robodebt case.

4.2.Review

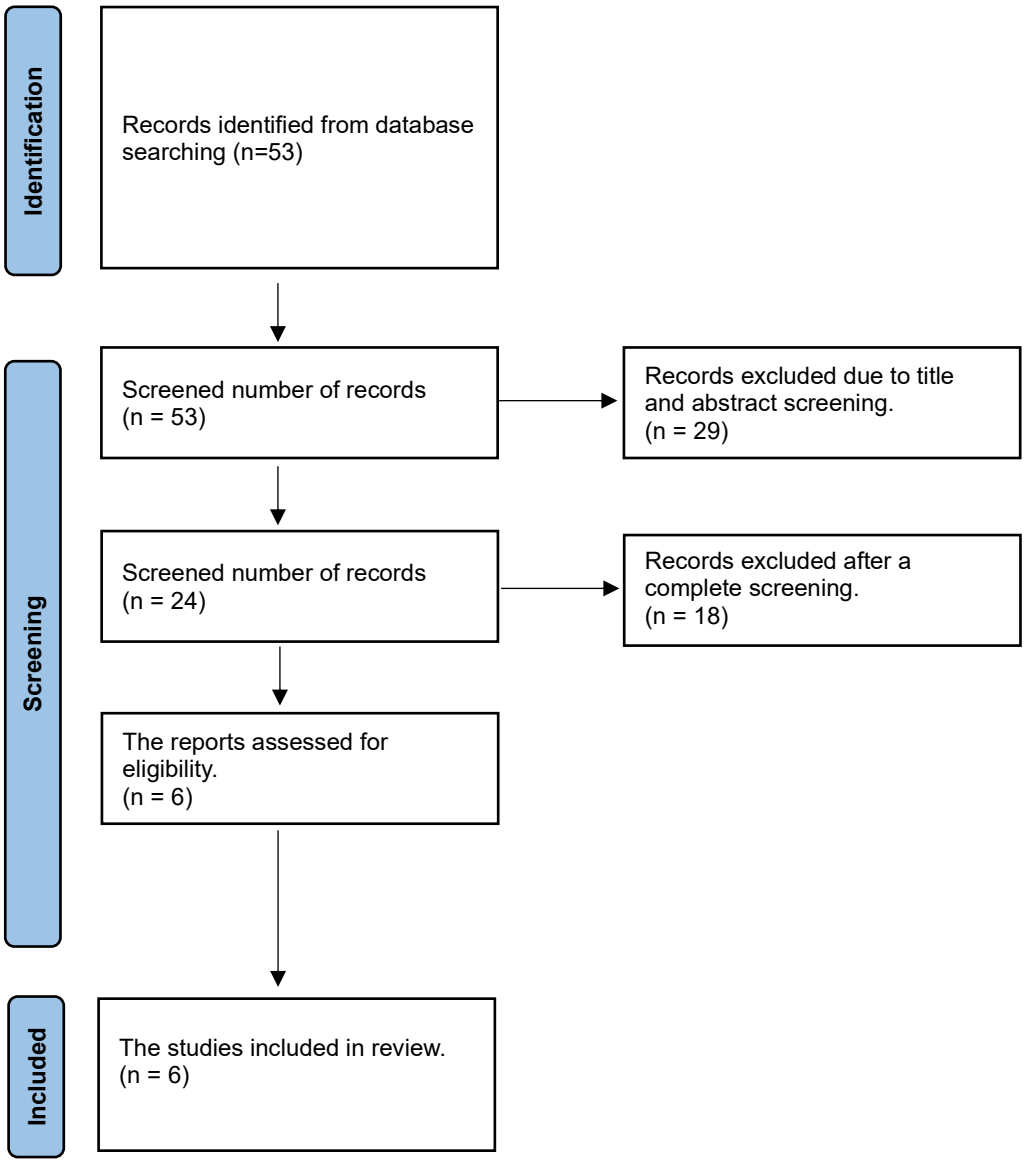
AI has been playing an active role in various fields for many years and it also sparks debates regarding their impacts. AI is not a novel concept, and it has been employed with different aims in different sectors. The existence of diverse fields of application may result in a range of definitions. Nevertheless, this study embarks on the review with the definition of European Commission. According to the European Commission (2018, p.1), AI is a "*system that displays intelligent behavior by analyzing their environment and taking actions – with some degree of autonomy – to*

achieve specific goals". In this manner, AI, machine learning or automated decision-making terms may be employed interchangeably.

As a first step of PRISMA approach for a literature review, a Scopus database is determined as a starting point, further, keywords and limitations that will be applied through document search also arranged. Scopus database is chosen to include as many articles as possible since it covers a wide range of journal articles. There are also other databases addressing artificial intelligence that might be added to this literature review. Given that the subject of this literature review is interdisciplinary, it is better not to use databases that are too specific and lean more towards the main technical aspects.

Using the specified keywords "artificial intelligence," "government," and "implementation," the search process is initiated. Additionally, the subject area was narrowed down to the domain of social science, and parameters were set to include articles sourced exclusively from English-language journals, particularly those accessible via open access. Following these initial criteria, a total of 53 articles were successfully identified. Later, the screening process took place based on the title and abstract. 19 articles are eliminated due to title screening and later, 10 articles are excluded after abstract reading. Having read all the remaining studies, I ended up with 6 articles around which I construct this literature review. No duplicate records or exclusions exist beyond those already specified.

In the end, six articles were identified through a thorough review process, comprising works by Ferretti (2022), Hinton (2023), Rachovitsa & Johann (2022), Drake et al. (2022), Chong et al. (2022), and Broomfield & Reutter (2021). These articles were carefully selected to serve as the foundation for the subsequent examination of the utilization of fairness within the realm of artificial intelligence in governmental contexts. Their inclusion in this study reflects a concerted effort to draw upon a diverse range of perspectives and insights from contemporary scholarship, thereby enriching the depth and breadth of the analysis.



Main attention in the literature seemed to be given to how AI related implementation in public institutions take place, however, other facets such as legal aspects or technological advancements are also covered along with ethical considerations. The issue of fairness is addressed in line with ethical concerns of AI implementation. While the concept of fairness is not directly discussed in the literature regarding AI implementation in public institutions, ethical approaches are addressed in a pretty similar manner. This draws attention to issues such as transparency, legal liabilities and discrimination while most of them are not empirical studies. The current literature is mainly shaped around two forms of justice: distributive justice and mostly procedural justice. The studies are not

directly talking about how these justice types are addressed, rather, what they are studying is related to one of these justice types.

4.2.1. Distributive Justice

Although based on complex algorithms created by highly skilled individuals, algorithmic decision-making systems impact every aspect of the lives of ordinary citizens. It is an important issue how these systems participate in the daily practices of social life and whether all stakeholders of society have equal rights and opportunities in this participation process. In this manner, literature suggests that AI implementation in public institutions is one of the ways that distributive justice takes place. Distributive justice is a concept of justice that ensures that resources are distributed fairly in a society and that everyone receives the share they deserve. This type of justice emphasizes that income, wealth, opportunities, and other resources should be distributed equally or fairly. In the context of algorithmic decision-making, distributive justice refers to the effort to ensure that the distribution of resources associated with the use of algorithms is fair and equitable. This involves analyzing how income, opportunities and other resources are distributed, while assessing the impact of algorithms on different groups within society. Complying with the principles of distributive justice in the use of algorithms aims to ensure that decision-making processes occur in a fair and balanced manner.

While the government is typically assumed to bear the responsibility of ensuring fair outcomes, private businesses also hold a share of this responsibility (Ferretti, 2022). In addition to governments establishing AI ethics guidelines through a democratic and legitimate process, private companies also play a role in contributing to the development of a legal framework that prioritizes ethics in the absence of explicit legal guidelines. This way it can be ensured that those who are affected by the distribution of costs and rewards are participating in the process. Similarly, Hinton (2023) draws attention to the importance of cooperation of the relevant parties. During the development stage, implementors of AI solutions should proactively incorporate principles to mitigate the risks of unintended harm. Additionally, even indirect stakeholders like citizens should participate in the design process of AI systems involved in public service delivery or public interaction, as their usage directly affects them. Input from these indirect stakeholders can potentially resolve unfair tendencies within the algorithm even before putting it into service and help produce fair outcomes. The equitable distribution of resources in contexts where algorithms

are employed to make decisions on behalf of interested parties can be ensured through the participation of potentially affected stakeholders.

4.2.2. Procedural Justice

The majority of the literature concerns procedural justice. It emphasizes the significance of not only the outcome but also the procedural aspects of decision-making. Procedural justice refers to the fair conduct of the decision-making processes of an institution or system. This type of justice aims to ensure that decision-making processes are transparent, impartial, participatory, and consistent. How people are treated during their interactions with the system affects their perceptions of fairness in that system, to some extent, regardless of the consequences.

In the context of algorithmic decision-making, procedural justice refers to the fair and transparent conduct of decision-making processes using algorithms. This includes efforts to understand and evaluate how algorithms make their decisions, what data they are based on, and how those decisions are implemented. Procedural justice principles aim to ensure that decision-making processes occur in an impartial and consistent manner, while encouraging participation, accountability, and transparency in the use of algorithms.

In relation to algorithmic decision making in public institutions, Rachovitsa & Johann (2022) discusses about human rights implications of AI, based on the Dutch Syri Case, which is also one of the cases of this study. They claim that such a flawed implementation causes three main adverse effects in human rights; law, data protection, and algorithmic accountability/transparency. A lack of transparency in implementation of algorithmic systems is not only a determining factor over effectiveness but also a determining factor over adequate judicial supervision. When openness of the system is a matter of debate then it becomes difficult to ensure fairness both within the system and its consequences. Similarly, increasing accountability and dialogue activities will also improve the quality of the participant experience in the algorithmic decision-making context. Especially if regular information is provided about the process, this will increase the awareness and knowledge of ordinary citizens about the system (Drake et al., 2022).

A study draws attention to the procedure itself to gain the maximum benefits in using AI in urban technology. It is argued that facilitating collaboration with open data and establishing a framework that everyone agrees on will increase efficiency in the context of urban mobility (Chong et al., 2022). In this way, expectations regarding the outcome are agreed upon. When it is known what

the data contains, it is possible for the outcome to be fair. The process definitely affects the outcome, either in terms of fairness or in terms of the quality of the service.

Having said that, putting data at the center does not always come as easy. From organizational culture, technical capabilities to regulatory frameworks, many things need to be considered when a data driven approach is intended (Broomfield & Reutter, 2021). In the employment of AI by public institutions, the process is as effective as the result. It is important to be accountable and transparent to ensure procedural justice. At the same time, given the importance of data, system design is also very important both technically and administratively.

Overall, the literature examining artificial intelligence implementation in public institutions from the perspective of fairness is scarce. Those who touch upon this area are generally taking this issue from a procedural and distributive justice perspective. While it is recommended in the literature that relevant initiatives include society in the process, the importance of transparency and accountability is also highlighted. The issue of algorithmic fairness has also been examined concerning machine learning, but those studies have been excluded because they fall outside the social sciences and focus only on the technical possibilities and capabilities of the process.

4.3.Conceptualization of Fairness in relation to Algorithmic Decision-Making

Algorithmic decision-making is the process that takes place when data are gathered through digital devices or systems and later processed by algorithms to make decisions (Newell & Marabelli, 2020). According to Sueur (2015, p.2), "*Automated decision-making involves breaking down a decision into a set of 'if-then' rules and criteria: a decision is understood as an algorithm (a sequence of reasoning) that selects from predetermined alternatives.*" Algorithmic decision-making processes impact many fields like education, employment, health care, credit, and government services (Monteith & Glenn, 2016). However, this procedure may come with some unintended consequences. Kuziemski & Misuraca (2020) claims that, in the context of the public sector, the desire to enhance efficiency conflicts with safeguarding citizens from potential algorithmic risks. Despite the efficiency algorithms offer, heavily relying on them in decision-making processes presents significant challenges ensuring equitable outcomes (Köchling & Wehner, 2020).

Leventhal (1980) describes fairness as equal treatment based on people's performance. In this manner, fairness is closely tied to individuals' activities rather than their identity or origin.

Leventhal also discusses that rewards or punishments need to originate from an act. In this context, algorithms lead to a process in which a person's performance becomes unimportant. They make decisions about the future based on data that may not directly reflect people's performance. The decision that concerns the person is taken independently of the person's behavior regarding that issue.

Fairness can also be described as the lack of discrimination in the process of making decisions (Farnadi, Babaki, & Getoor, 2018) while the main expectations are for AI systems to maintain fairness in terms of respecting the law, human rights, and democratic values and principles (Xivuri & Twinomurinzi, 2021). Fairness necessitates that AI is developed in a way that advances freedom and equality (Ienca, 2019). This involves treating others without any personal bias or self-interest, both in the outcomes and the processes leading to those outcomes. From this point of view, it is also crucial to persuade people that the algorithmic decision-making process is built in a way that produces trustworthy outcomes if it is employed in public institutions. Designing algorithms fairly is of high importance to ensure public trust. The parliamentary Diana Baran, in a parliamentary debate, argues that it is essential to guarantee the preservation of public trust in the algorithmic decision-making processes concerning them and the utilization of their data, thereby upholding fairness (UK Parliament, 2020). To a certain degree, fairness is tied to perception. Building trust involves effectively communicating with people and ensuring they comprehend the algorithmic decision-making system. Complexity within the system should be reduced in the design phase. Wang, Harper, and Zhu (2020) et al claim that people's perception of an algorithm's fairness is affected by a system's clarity and openness. A certain degree of transparency regarding an algorithm tends to enhance users' trust in the system, even if their expectations don't match the algorithm's outcome. Still, excessive disclosure about the algorithm carries the potential risk of confusing users, thus diminishing the perceived fairness.

Algorithmic fairness is recognized as one of the four primary principles for establishing trustworthy Artificial Intelligence (AI) by the OECD (2019). The European Parliament (2018) contends that debates regarding fairness in algorithmic decision-making systems frequently lack rigor and precision, often relying on rhetoric. Defining fairness is a complex task, with numerous proposed definitions, some of which may be incompatible with each other. Thus, breaking down fairness into measurable phenomenon to extrapolate meaningful insight into algorithmic decision-making is of

high importance. It is also important to make sure that these systems do not produce output that might be deemed discriminatory by a certain group of people within a society. However, it is difficult to claim a consensus on how to precisely define fairness of predictive models and how to assess their performance in terms of non-discrimination (Žliobaitė, 2017). While there are statistical or math-based approaches to define fairness in algorithms, the aim is to enhance public safety while adhering to non-discriminatory approach (Corbett-Davies et al., 2017). Statistical definitions of fairness categorize individuals into "protected groups" (frequently determined by factors like race, gender, or other binary protected attributes) and aim to balance certain classifier statistics (such as error rate, false positive rate, positive classification rate, etc.) across those groups (Sharifi-Malvajerdi, Kearns, & Roth, 2019).

The process of algorithmic decision-making should neither lead to situations where individuals experience injustice based on factors such as race, ethnicity, social class, or financial status nor prevent someone from enjoying his rights. It is imperative that algorithmic systems are designed and implemented with careful consideration for the principles of equity, fairness, and human rights. Furthermore, the potential biases embedded within algorithms must be continuously monitored and addressed to ensure that discriminatory outcomes are mitigated. Additionally, transparency and accountability mechanisms should be integral components of algorithmic decision-making processes, allowing for scrutiny and redress in cases of perceived injustices. By upholding these principles and fostering a culture of ethical algorithmic governance, we can strive towards a society where algorithmic systems contribute positively to equitable outcomes and uphold the rights and dignity of all individuals.

While literature suggests that the perception of people plays a role in whether they feel unfair it is better to design algorithmic decision-making systems transparent, simple, participatory, and through a legitimate process.

5. Research Findings

In relation to algorithmic decision-making systems, distributive justice and procedural justice came into prominence in our literature review. There while, in the theory section, Rawls's approach to justice also justifies the rationale to create codebook in a way to prioritize and seek these justice types in the existing textual data. Rawls's difference principle operates in a similar way given that distributive justice is about the equitable distribution of resources and opportunities in a society. In this principle, Rawls is concerned with the distribution of economic and social opportunities in a society and argues that these economic and social inequalities ought to be regulated according to the benefits of the least advantaged segment of society, which clearly speaks of distributive justice. In relation to that, the principle of equal liberty states that every individual should have an equal right to the broadest fundamental freedoms, consistent with the freedoms of others. In short, it emphasizes the protection and equal distribution of individuals' fundamental rights and freedoms.

On the other hand, the concepts of the original position and the veil of ignorance that Rawl proposed also indicate the procedural justice. Procedural justice concerns the process and procedures that produce results. The original position and the veil of ignorance concepts are crucial to produce fair outcomes by ensuring that the decision-making process is exempt from any biases or inclinations. Overall, the theoretical framework and the literature review of this study informed the creation of the coding scheme. While transparency prioritizes procedural aspects of algorithmic systems oversight concerns deals with distributive justice of algorithmic systems. Additionally, the importance given to transparency in the auditing of algorithmic systems emphasizes the importance of ensuring accountability and justice in the decision-making process. Moreover, the necessity of human intervention functions in response to distributive justice concerns by providing the opportunity to correct systemic biases and inequalities.

The initial coding scheme is presented as follows.

Table 1: Initial Coding Scheme

Theme	Sub-Theme	Codes	When to Use (When…)
Transparency	Communication Problems	Ineffective Engagement with Public	the references criticize public relations
		Inadequate Disclosure	criticism is expressed about insufficient information about how algorithms work
	Lack of Clarity	Ambiguous Structure	The structure of the whole system is not explained clearly.
		Unclear Direction	there are criticisms about how to object to decisions or taking them to the next authority
Oversight Concerns	Lack of Human Oversight	Insufficient staffing	there are not enough personnel to audit the algorithms or detect errors.
		Informational Void	there is no clear information about the whole process
	Complexity of the System	Complicated Procedures	there are criticisms that the system is difficult to understand or monitor.
		Inadequate Safeguards within the system	the lack of control mechanisms within the system exists

(Source: author)

5.1.Robodebt

The Australian Online Compliance Intervention Program which is known as Robodebt case refers to the automated decision-making scandal that occurred in Australia between 2016 and 2020. Centrelink is an Australian government agency responsible for delivering a range of welfare payments and services, including income support. The Robodebt program was operated under Centrelink's umbrella, specifically within the context of welfare payments and debt recovery. This system was launched as a budgetary savings program aimed at reducing the financial gap. However, the expected results were not achieved during the implementation process and the performance of the system was unexpectedly poor. This situation demonstrates the difficulties and ineffectiveness of the system in achieving its initial goals. Such a scenario, which causes significant damage not only financially but also in terms of humanity and ethical values, also negatively affects the mental health of the society and causes deep destruction. In this context, it is seen that the effects of the system are wide-ranging and in various dimensions.

The period spanning from 2015 to 2019 witnessed the implementation of the Robodebt scheme, officially designated as an income compliance program. This initiative primarily functioned by employing an algorithmic framework to cross-reference individuals' reported incomes with the corresponding social welfare benefits they received, a mechanism commonly referred to as income-averaging. Discrepancies detected between reported income and welfare entitlements underwent scrutiny, with recipients being subject to potential repayment obligations commensurate with the degree of perceived overpayment. In 2015, the coalition government launched this program to recoup overpayments, aiming to detect cheaters within the social welfare system and address budgetary deficits. A sum of AUD $746 million was mistakenly collected from 381,000 individuals and subsequently reimbursed. Similar to SyRI, the implementers of this program also operated under the assumption that a certain stratum of society has the potential to exploit the system, in this case, social welfare recipients.

Australia has demonstrated a longstanding interest in welfare fraud detection, dating back to the 1990s with the enactment of the Data-Matching Program (Assistance and Tax) law. This legislative measure aimed to enhance welfare program integrity through the implementation of data matching systems. While the effectiveness of these efforts varied, they achieved measurable success to some degree. Until 2015, the data matching system was carried out manually by officers by manually matching an individual's annual income reported to Centrelink with relevant data in the Australian

Taxation Office's (ATO) records. However, the Income Compliance Program came into effect in 2015, and with this program, the responsibility for explaining any discrepancies detected was transferred to individuals. However, it was later found that the Robodebt algorithm further complicated the issue.

The fundamental flaw of the data matching system was that the data was analyzed over different time periods by two different agencies, the ATO and Centrelink. While the ATO assessed annual income, Centrelink was based on fortnightly income. This was based on the condition that individuals should not change workplaces during the year, but this condition was not enforceable for many individuals. Herein lies the fundamental systemic problem that causes the income averaging system to fail. Another problem that arises from this logical error is the need to prove that you correctly declared two weeks' income a few years ago. Automating the system without making any improvements to these very obvious errors of the income averaging system has only created a snowball effect in terms of negative effects.

While this system appears sensible in a general sense, its underlying assumptions were susceptible to distortion, undermining its primary intent. The main flaw in their assumption lay in the expectation that individuals subject to scrutiny possessed a consistent income over a specified timeframe. However, this assumption fails to hold true for everyone; some individuals may experience prolonged periods of unemployment or frequent transition between temporary positions. Consequently, compiling an accurate income history becomes challenging. Furthermore, victims of the system are required to substantiate their claims when contesting erroneous fines, thereby shifting the burden of proof onto them. Given that the overarching objectives of digitizing social systems encompassed enhancing effectiveness and efficiency, this approach ultimately failed to meet its intended goals and instead exacerbated hardships for certain individuals. It merely transferred the burden from one party to another, in this case, the citizenry. It soon became evident that the entire system was flawed, prompting a surge in public discontent with each passing day.

The systemic irregularities and lack of rigor, both inherent within the system itself and among its implementors, are also echoed in the e-petitions submitted by individuals. Within these petitions, people commonly expressed the view that the process was unfair, yet they approached the issue of fairness from various angles and perspectives.

At the outset of the process, participants were already of the view that no convincing argument had been presented regarding the debt alleged against them. The problem in terms of transparency and accountability is not the issues that arise as the process progresses, but rather the reasons that lead to the process being so problematic. In fact, the situation that emerged at the beginning of the process persisted not during the process itself, but in the subsequent investigation process. For instance, victims' personal information was shared with the media, with the aim of deterring them from speaking out further (Jones, 2023). As in Syri, government officials did not give much attention to the confidentiality of private life. Another detrimental mistake is that some senior level public servants misled the cabinet by submitting a partial document (Utting, 2023).

Transparency, when it comes to automated decision making, constitutes one of the core issues that needs to be considered. The Automated Decision-making Better Practice Guide that is published by the Commonwealth Ombudsman states that “Automated systems must comply with administrative law principles of legality, fairness, rationality and transparency”. The guide also specifies that transparency and accountability need special care when designing and implementing automated decision-making systems. This is a commonplace occurrence that algorithm-based systems suffer from a lack of transparency, rendering their structural logic difficult to comprehend. It is the main case for Robodebt. Many people didn’t even know why they were in debt or how should proceed or avoid payment. This is mainly due to the opaque structure of the Robodebt system.

It is an important contradiction that Centrelink sends notifications to people that they owe money and then holds them responsible for proving the contrary, because until the end of 2016, on their website, they advised them to keep their pay slips for only six months. Given the pattern of people's record-keeping behavior in the context of Centrelink for six months, it was not reasonable to expect several years' worth of records (Carney, 2018). Still, these people could be considered lucky because some people did not receive any notification that they owed money and needed to prove if they thought there was a misunderstanding. Their first interaction was with debt collectors to whom Centrelink outsources debt collection. In such a situation, some people who wanted to contact Centrelink either had to endure a hold-on period that could last for hours or had to overcome the indifference of the employees. These errors in the system did not prevent Centrelink from adding a 10 percent penalty to the debts (Marks, 2020).

At the same time that the Robodebt application was launched, federal government elections were held in Australia. This election was highly competitive, and the Liberal-National Coalition government was formed. It should be debated that a mistake in the design of this program is that it was arranged to be brought into force in the atmosphere of the election period. Politicians may make various electoral commitments and wish to achieve results in order to fulfill them quickly. Moreover, considering that the most important agenda during the election period is to win the election, sufficient attention may not have been paid to such practices.

One of the issues that makes Robodebt interesting is that this fiasco is not directly related to the incompetence of the Australian government. Luetjens, Mintrom, & Hart (2019) present the public policies successfully implemented by the Australian government in their study titled Successful "Public Policy: Lessons from Australia and New Zealand". This shows that the Australian government has the capacity to implement policy changes that concern individual citizens effectively. Clearly, this positive potential of the Australian government emphasizes the extraordinary nature of this failure. In particular, despite these successes of the government, encountering a faulty application such as Robodebt caused a great reaction and criticism from the public. This situation also reveals the need for a more careful evaluation of the government's policy practices. It also shows that when the issue is artificial intelligence or algorithmic decision-making systems, conventional public administration approaches may not work, and the process should be managed by civil servants who are familiar with the nature of the issue.

Table 2: Coding Frame[1]

Theme	Subthemes	Definition	Example
Psychological Well-Being	Mental Health Problems	This subtheme includes a variety of psychological disorders that affect an individual's mental health. For example, it includes conditions such as depression, anxiety, or bipolar disorder.	Nevertheless, I found this entire process of having to disprove a false robodebt exhausting, a waste of mine and others' time, and deleterious to my mental health (Anwyn Crawford ANON-24KG-9SC6-W)
	Emotional Damage	This subtheme includes traumatic experiences or negative emotional states that affect an individual's emotional health. For example, it covers situations such as losses, traumas or relationship problems.	I am now totally filthy with most things in life I have become a recluse I live alone hate 99.9 percent of people. (Trevor Windle ANON-24KG-9S1C-R)
Political Awareness	Development of Politically Reactive Behavior	This subtheme examines the development of individuals' reactive behaviors towards political processes.	I hold strong resentment to the inaction from the many politicians I contacted in early 2017. … How can it be so complicated? 'The government is running a speculative invoicing scam, like the type your grandma got done by, but it's the government and you HAVE to pay.' It's not that complicated, surely? (Michael

[1] Examples provided both herein and in the subsequent sections can be obtained from the submission website at https://robodebt.royalcommission.gov.au/submission.

			Griffin ANON-24KG-9B1E-9)
	Feeling Tends to be Criminal.	This subtheme includes individuals' feelings of distrust or pressure towards political processes	"I hold strong resentment to the inaction from the many politicians I contacted in early 2017. (Michael Griffin ANON-24KG-9B1E-9)
Transparency	Communication Problems	This subtheme includes communication deficiencies and problems within the institution or process. For example, it covers problems such as breaks in the flow of information, misunderstandings, or communication barriers.	Communication from centerlink themselves was very poor. (Sarah Eastwood ANON-24KG-9SBT-T)
	Lack of Clarity	This subtheme examines uncertainties and unclear situations in the institution or process. For example, it includes situations where policies or guidelines are unclear, decision-making processes are unclear, etc	I was sent reams of paper that made no sense... I still did not agree but paid a minimum amount each week. (Graeme B ANON-24KG-9SG2-W)
Oversight Concerns	Lack of Human Oversight	This subtheme includes problems related to the lack of human control of a process or system. For example, it covers issues such as overuse of automation, inadequate human supervision, or reduced human intervention.	A simple human check would have shown this was not possible. (Graeme B ANON-24KG-9SG2-W)

(Source: author)

The examination of how the issue of fairness is addressed around three cases has been conducted through textual data comprising petitions, media reports, and academic studies. A thematic analysis

method is employed to identify patterns within textual data. The decision not to use quantitative analysis software in this study was based on the feasibility of conducting the analysis given the amount of available data. One hundred and twenty e-petitions are analyzed. Every one of them was read thoroughly and relevant sections were highlighted in accordance with the initial coding scheme. Later, codes are assigned based on these highlighted sections. This process allowed documents to be re-read and checked.

There is no hierarchical relationship between the created categories. Each category was used to group codes with similar characteristics. These categories represent different aspects of the research and are each independent of each other.

Emerging themes out of the e-petitions are “psychological well-being”, “transparency and accountability”, “oversight concerns”, and “political awareness”. While the subsequent themes predominantly pertain to individuals' perceptions of fairness, the initial theme does not directly engage with fairness issues. Nonetheless, due to both the fact that it is in relation to other themes and its conspicuous presence within the corpus of e-petitions under examination, it has been included in this study as a suggestive avenue for potential research exploration.

Over 800 individuals reportedly died by suicide subsequent to receiving debt notifications. Among these, 550 had made at least one payment prior to their deaths. While direct causality between these deaths and Robodebt notifications has not been established through empirical study, the significant psychological distress induced by sudden debt imposition is evident. The Australian government has maintained that correlation does not imply causation, emphasizing the necessity for additional evidential support (Sally, 2019).

However, the Royal Australian and New Zealand College of Psychiatrists (RANZCP) claims that the act of labeling individuals as cheaters could potentially result in trauma, anxiety, distress, and mental health complications (RANZCP, 2023).

“For many of the victims there is no satisfactory remedy. Their lives were impacted so greatly and, of course, the people who self-harmed and suicided, those families will never get their loved one back. Nothing will ever make this fully right” (Karp, 2023). Graeme B expresses that “The timing of Centrelink and dealing with the stress really pushed me to thoughts of suicide... I was lucky."

Even though the themes are not in hierarchical order, however, they are still interrelated. The other themes may cause mental health problems indirectly.

Secondly there are the transparency and accountability themes. Transparency and accountability are related to each other and the existence of the one lead to another one. Transparency ensures that an institution or process is accountable because the activities of a transparent institution can be more easily monitored and evaluated. In addition, accountability can arise as a result of transparency; That is, when an organization clearly articulates its activities and communicates with stakeholders, it can increase its sense of accountability. It can be claimed that Robodebt is failed in this manner and this aspect of the story is very well articulated in the petitions.

Some petitioners express the lack of transparency and accountability by highlighting the communication problems. The debt calculation is unclear, and there was insufficient effort to reassure people about the safety of the system. Even when they have a chance to get in touch with the person of interest things get more complicated. Daniel B states "*2017, the debt was raised again, and it was a different amount... 2018, Centrelink raised a new debt, a different figure...*"

Another category that emerged in the petitions is oversight concerns. This theme encompasses both lack of human assistance when people feel a need for a consultation, and system's complexities. People voiced their concerns about the challenge of reaching a key individual. McMahon states that it took hours for him and was placed on hold just to be able to speak with a person. Furthermore, he contends that this lack of response further exacerbated adverse effects on his mental well-being. Another individual affected, Karen Ryall, also contends with a comparable issue. Expressing dissatisfaction with the challenge of finding personnel to oversee the process, Ryall asserts that this predicament is particularly exacerbated when operating from a foreign jurisdiction -overseas-.

Lack of human assistance is felt a lot, mostly due to the complexity of the system. While people are not aware of the underlying logic of algorithms, they didn't even understand how these debts were calculated. Not only was the system utilized by Centrelink complex, but the verification process it demanded was also comparatively intricate. One of the affected victims, Graeme B, highlights the intricacy involved in gathering salary records from the various workplaces he has been employed at in the past, "To challenge this I had to collect pay slips for all 8 employers... and supply to Centrelink which was arduous."

The last theme surfacing the petitions is "political awareness," which implies that individuals who have been subjected to unjust treatment by the state and labeled as fraudsters develop reactive political behaviors. Above all, people have lost their trust in the government, at the very least. The increasing loss of trust in the government over time has led individuals to develop political reactionary attitudes. Within this context, one reason for their stance is their "oversight concerns," meaning they cannot find individuals to address their grievances or articulate their concerns to, coupled with the lack of transparency in the system they interact with. Consequently, they direct their frustration towards the government.

Michael Griffin, a victim of the situation, articulated his sentiments as follows: "I hold strong resentment to the inaction from the many politicians I contacted in early 2017." Many politicians failed to acknowledge or critically consider the possibility of flaws within the system. This misconception resulted in a scenario where victims were unjustly perceived as deceivers. Griffin further remarked, "How can it be so complicated? 'The government is running a speculative invoicing scam, like the type of your grandma got done by, but it's the government and you HAVE to pay.' It's not that complicated, surely?". Another victim, Peter Dickson, pointed out that "Such politicians also have a view that those who are receiving benefits are a burden on society," highlighting politicians' perception of themselves as a burden on society. Robodebt victims organized street protests in various locations across Australia.

For the remaining two cases of the study, our data consists of media reports and academic studies obtained through LexisNexis. LexisNexis is a Global Information & Analytics Company, and it provides legal, regulatory and business information and analytics (LexisNexis, 2024). In this context, while 100 material was taken for the SyRI case, 70 material was accessed for the A Level Exam Case. Moreover, the analysis shows how the issue, contrary to the Robodebt case, is discussed in the public sphere and which features of the issue are highlighted compared to the first-degree victims of the incident.

5.2.SyRI

Systeem Risico Indicatie (SyRI), akin to Robodebt, is a digital welfare fraud detection system. Such endeavors are intrinsically linked to the burgeoning trend of digitalization. The COVID-19 pandemic has underscored the extent to which individuals depend on social security systems and their susceptibility to transitioning to such assistance programs. This underscores the imperative

for the digitization of welfare systems. Consequently, the efficacy and functionality of social security systems has become paramount. The process of digitizing welfare systems aligns with the overarching goal of ensuring their efficiency and effectiveness.

Although it is similar to Robodebt in terms of being a social welfare fraud detection system, there is an important difference between them. Robodebt identified inconsistencies in the data so that the existing output, even if inaccurate, was a solid one. However, as a result of processing the huge amount of data it collects, SyRI identifies profiles that are worth examining - those that are likely to deceive the system. What exacerbates the issue further is that this entire system is integrated into the Dutch Government's neighborhood-focused approach aimed at ameliorating conditions within disadvantaged communities. These communities are known for their crime rates, and poverty; the SyRI system is employed mainly in these areas, not the whole country. Considering that SyRI is applied in certain regions, it is understood that some people are automatically assumed as potential criminals.

One significant aspect highlighting the importance of SyRI is its cessation by the court due to privacy concerns and human rights considerations. It was found to be contrary to ECHR article 8 rather than the European Data Protection Framework of EU (GDPR). GDPR gives individuals the right to opt out of fully automated decision-making processes in situations that may have legal consequences. However, if there is a human anywhere in the process, this can make the process more complicated. In this context, in order to prevent similar discussions altogether, the court addressed the decision in the context of human rights by referring to the ECHR, and the decision is a decision that may affect not only the SyRI context but all similar systems. Article 8 mandates that all legislation maintains a "fair balance" between societal interests and potential infringement upon citizens' private lives. The objective of SyRI, defined as a "social interest," is to detect and combat fraud. While the Dutch government argued that the SyRI legislation adequately safeguards privacy while achieving its goals, the court held a different view that the legislation lacks transparency and fails to provide adequate protection against privacy violations.

There were also two similar algorithm-based systems in the Netherlands, which are Waterproof (2004-2007) and Black Box (2008-2017) (Wieringa, 2023). Therefore, the SyRI practice did not arise abruptly; rather, its origins can be traced to these antecedents. SyRI passed both the House of Representatives and the Senate without any debate (Wieringa, 2023). The fact that there is no

discussion about SyrI in these two institutions does not mean that the issue is not discussed at all. On the contrary, the issue has brought about many discussions. In 2014, before it came into force, it was heavily criticized by both the Data Protection Authority and the Council State. SyRI is not an unknown creature that suddenly came from outer space.

The Dutch government uses a risk modeling algorithm, the details of which it does not disclose, to classify its citizens. The main goal of the SyRI was to detect potential frauds and lead the way for further investigation. Although it is generally seen as a logical and socially beneficial initiative, the questions of which data are handled on an individual basis and how they are processed are not clear. Moreover, the exclusive implementation of SyRI in impoverished neighborhoods or areas deemed by the state to have a higher likelihood of fraud elucidates why it was the first system to be revoked in the context of human rights.

Although SyRI is not discussed thoroughly by the two necessary institutions mentioned above, it has brought about many discussions in the public sphere. In 2018, Platform Bescherming Burgerrechten, together with some Dutch non-governmental organizations, filed a lawsuit against the government to stop SyRI. With this legal process they initiated, they also became more visible in the media with the campaign they ran and ignited a public debate (AlgorithmWatch, 2020).

Three primary categories have been identified from the analyzed media and academic documents. Each of these categories comprises subcategories as shown in the table above. These categories are moral implication and discrimination, privacy concerns, transparency, and accountability.

The moral implication and discrimination category covers fundamental concerns facing many social and technological applications. Under this category, public discourse appears to revolve around the ethical implications of SyRI, delineated into two subcategories: ethical consideration and unfair targeting. The first subcategory deals with the uncertainties or difficulties experienced at the ethical level about the rightness or wrongness of an the SyRI. Human rights occupy a place at the center of the SyRI debate. This atmosphere enabled ethical questions to be visible in public discussions. On the other hand, rather than ethical problems embedded in the system it also directly targeted those who might have that characteristics expecting that these people are more likely to cheat the system. Targeting underdeveloped neighbors is part of a neighborhood-oriented projects or wijkgerichte aanpak in Dutch. In this manner, targeting low income neighborhoods are not unique to SyRI, there are other instances showing that these areas are specifically brought under

the lights. Documents reveal that while such systems can produce effective results, at the same time, they contain a morally controversial problem and especially some groups of people are more likely to experience unfair targeting. Considering SyRI's deployment in regions inhabited by socioeconomically disadvantaged households, one could anticipate robust discourse surrounding this facet of the issue.

The second theme in the documents pertains to privacy apprehensions. The Privacy Concerns category includes concerns about protecting personal data and ensuring privacy. This category has two subcategories which are privacy violation and data protection concerns. Privacy Violation covers situations where people's personal information is captured or disclosed without permission. Data Protection Concerns refers to concerns about inadequate protection of personal data. The ambiguity of the SyRI case raised concerns about people's privacy. As will be elucidated in subsequent chapters, the algorithm's structure remains undisclosed to the public, with the government withholding it even from the court. This opacity has engendered concerns regarding privacy among the populace, a sentiment mirrored in ongoing discussions. The government believes that any attempt to reveal how SyRI works will make a welfare system more vulnerable against fraudsters (Human Rights Watch, 2019). The government failed in securing the balance between the national security and the privacy right of individuals. In this case we see that the government prioritized national security.

This theme points to an important issue in the context of the implementation of algorithmic decision-making: opacity in decision making process. The documents examined under this category include details such as how algorithmic systems work, what data is used, and which algorithms are preferred. These details are important for understanding and publicly explaining how systems' decisions are made. In this way, it can be ensured that systems are transparent and reliable, and public institutions can be held accountable for these decisions.

Accountability/transparency and oversight concerns are taken into consideration together in line with the judicial process in the public debates. This caused the SyRI case to relatively deviate from the same concerns of Australian people. The distinction arises from the fact that in the Robodebt case, e-petitions were predominantly initiated by individuals directly impacted by the algorithm, while in this case, discourse involves a broader spectrum of stakeholders. Additionally, Robodebt

petitions were lodged prior to judicial rulings, whereas the present materials predominantly pertain to the post-judgment period.

The initial subcategory revolves around diverse methods of ensuring accountability while concurrently upholding the primary purpose or significance. Several documents scrutinize the SyRI algorithm, highlighting its opaque framework and insufficient mechanisms for accountability. Additionally, critiques extend to the accountability of non-algorithmic aspects within the narrative. The democratic prerequisites concerning SyRI were inadequately met owing to the government's position on these matters. In an environment where the government withholds the algorithmic framework from the media or judiciary, accountability extends beyond the algorithm itself.

In the secondary subcategory regarding oversight apprehensions, there exists a subtlety in the role of oversight. While it portrays a surveillance or intervention mechanism within the algorithmic system as either aiding implementation in e-petitions or as a means to attain resolution, the focus here extends to judicial oversight as well. It not only encompasses the procedural aspect but also addresses the aftermath within the realm of oversight apprehensions. Certain documents advocate for legal scrutiny to safeguard individuals' rights.

In summary, despite comprising three distinct categories and six subcategories, each is interconnected and capable of influencing the others. This interrelation is facilitated by the structural characteristics of algorithmic systems, while government attitudes can also contribute to this complexity. The topics under discussion are not isolated but encompass various facets of an interlinked public discourse.

5.3.A-Level Exams

The A Level grading scandal occurred in the UK in 2020, which resulted in many students receiving lower grades than expected due to an algorithmic error in the exam system. While the pandemic has indeed facilitated digital transformation across various sectors, it can be argued that the educational domain stands out as one of the most profoundly impacted by this paradigm shift. This repercussion was particularly pronounced in nations like the United Kingdom (UK). Notably, the UK made the decision to cancel all secondary school examinations in 2020 as part of its measures to address the challenges posed by the Covid-19 pandemic. The Advanced Level (A Level) and General Certificate of Secondary Education (GSCE) exams are held for secondary education students in their academic journey towards further educational opportunities, such as higher

education or vocational training. In a timely manner, in June, the Office of Qualifications and Examinations Regulation (Ofqual) developed a grade standardization algorithm at the request of Gavin Williamson, who held the position of Secretary of State for Education at that time. Williamson states that "*My priority now is to ensure no young person faces a barrier when it comes to moving on to the next stage of their lives – whether that's further or higher education, an apprenticeship or a job"* (Parker, 2020).

Ofqual is the regulatory body responsible for overseeing qualifications, exams, and assessments in England. Even though Ofqual was not an avid proponent of this algorithmic system they still provided the necessary system.

Ofqual, in their written statement, explains in detail their side of the process. When they were advised by the Secretary of the State regarding this issue it was their recommendation that the optimal course would be to continue with sitting exam by taking care of social distancing. However, moderating the grades determined by teachers was decided on. They also indicated that the grading of teachers itself is not standardized, while some teachers shows more generosity the others might not show similar performance. Moreover, socioeconomic differences might worsen socioeconomic inequality. From this point of view, the report states that introducing an algorithmic method looks pretty logical in principle, and also states that many education stakeholders have found it suitable to use statistical methods to eliminate potential side effects that might be caused by teachers' assessments. Still, they had some reservations. After all, such a system has the potential to disproportionately affect a small group and give rise to due process. They also describe such a task as impossible to achieve and conclude the statement with these remarks "*a 'better' algorithm would not have made the outcomes significantly more acceptable. The inherent limitations of the data and the nature of the process were what made it unacceptable.*" (Ofqual, 2020).

This algorithm is not designed to automatically grade students or place them without examination, but to prevent grade inflation in the grades awarded by teachers for A-Level and GCSE exams.

The system designed by Ofqual is generally based on information from teachers, which includes an estimated grade for a student, as well as the student's position relative to other students at the school. This is where algorithm steps in. In the light of this information, the algorithm would give relevant grades to the students. Its main purpose was to ensure that the grades given that year were compatible with the grades given in the past.

The logic was that teachers could be very generous in the grades they gave to their students, and this could be unfair to some students in the general sense. However, the account at home didn't match at the market. Upon the release of A-level grades in England, Wales, and Northern Ireland on August 13th, nearly 40% of grades fell below teachers' assessments. In England, 36% received a grade lower than teacher recommendations, with an additional 3% falling two grades below. Furthermore, the impact of downgrading was more pronounced in public schools compared to the private sector (BBC, 2020). Given that the main logic of Ofqual was to reflect last year's results, however, this meant that the grades of students who were successful above the school average were brought down to the school level. Likewise, schools that aimed to increase their overall success with the reforms they made were being sent to where they belonged, so to speak. On the other hand, private schools are schools that have achieved above-average success for many years because they have wider opportunities and admit students after an elite selection process compared to public schools. When it comes to algorithms based on past records, students studying in private schools will automatically have an advantage.

When looking for an answer to the question of why this algorithm was developed, one of the main motivations is related to the COVID-19 pandemic. It was predicted that university acceptance rates would decrease by approximately 24% due to the pandemic. This will lead to loss of income for universities and increase competition. In addition, eliminating the social risks of written exams within the framework of pandemic restrictions is another source of motivation.

Unlike the other two cases, the A Level algorithm worked with a different logic than the data matching method. There were three important points in the operation of the algorithm (Sky News, 2020). The first and perhaps the most important of these is the score distribution in the test center for that exam covering the years 2017 and 2019. The second input is the ranking of the relevant person in the test center. This sequence consists of two parts; The first is the person's ranking within the test center, and the other is the central evaluation score, which is roughly the teacher's predictive approach to the potential success of the relevant student. The last element is the person's past exam scores. This logic, to a certain extent, derived a person's success from the person himself and indexed it to the success of those who took that exam in the past years. If the person was in the middle ranks of the score distribution, it meant that he would receive a grade similar to the grades of people who were in the middle ranks in previous years.

After public outcry, the government backed down and students were graded based on their teachers' predictions of the grades they would achieve. However, with the cancellation of the algorithm, the grades of many students increased, which created a burden for universities. For example, Durham University has offered new students a variety of financial incentives, including scholarships and accommodation guarantees, if they postpone their registration this year until next year (BBC, 2020).

The A level algorithm has also brought to light the gap between private and government institutions. It is highlighted that the percentage increase in the number of top grades (A grades and above) at private schools was 4.7% annually, which was more than twice the rate observed at state schools, where it rose by 2% (Allegretti, 2020).

Through the analysis of related documents, this study delves into two overarching categories: Discriminatory Tendency, Accountability and Transparency. Within these categories, a nuanced exploration unfolds through subcategories such as Bias Concerns, Social Inequality, Criticism of Government Oversight, and Evaluation of Regulatory Leadership.

A-Level Exam case related documents draw attention to the discriminatory nature of the implemented system. The impact of the system is highlighted to be felt more by state schools' students since their opportunities compared to private school students in terms of education during the lockdown are more limited. This expectation arises from the potential biases embedded in the algorithm and the fear of exacerbation of existing social inequalities. Scotland that used a system alike for that period of time stepped back after it was realized that many disadvantaged students were downgraded (BBC News, 2020).

The second main category is transparency and accountability concerns. As well as other cases, this category shines out in A-Level Exam related debates. It is accompanied by two subcategories: Criticism of Government Oversight and Evaluation of Regulatory Leadership. In particular, criticisms have come to the fore that the methodology and reliability of the statistical model and its potential effects have not been adequately examined. Geofrey Barton, the general secretary of the Association of School and College Leaders, complains that the government has not established the necessary oversight mechanisms on this issue.

5.4.Synthesis of Case Learnings and Coding Scheme Enhancement

Up to this stage in the study, the Robodebt case has been examined as the main case, and SyRI and A Level case have also taken their place in the study as exploratory cases. An initial coding scheme was created at the beginning of the chapter and this scheme was kept in mind while examining the cases. In reviewing the Robodebt case, themes were also allowed to emerge from the data.

In this regard, the first coding scheme created based on theory and literature was designed to reveal themes related to both the results of the algorithmic decision-making system and the process of the relevant system in e-petitions. Although a direct coding process was not applied for the exploratory cases of this study, the initial coding scheme served as a guide in examining the cases.

The contributions of the cases, in addition to the initial coding scheme, are both the two themes added to the coding frame and the similarities of the cases to each other.

The essence of all the cases, in fact, the starting point of all of them and, in a sense, the point that makes the process unmanageable is the pursuit of efficiency. The assumption that such a pursuit is ultimately an absolute necessity in the context of algorithmic decision-making systems is the prevailing view in these cases. This assumption error is also, in a sense, the reason for the oversight problems experienced in the cases examined. For instance, in the Robodebt case, there was so much confidence that the system would work correctly and produce correct results that there was no need to install a control mechanism or intervene in the system.

The importance of the order among Rawls' principles is also relevant here. In Rawls' theory, the principle of equal liberty has priority over other principles. Similarly, in algorithmic decision-making systems, certain priorities must be established and maintained, ensuring that fundamental values are never compromised. These priorities should ensure that the algorithmic system is executed fairly and that its decisions are just. Based on this, the initial coding scheme has been designed to reflect procedural justice and distributive justice. In the study, the prepared initial coding scheme takes into account these considerations within the themes of transparency and oversight concerns.

As the research progressed, new insights emerged, which were subsequently incorporated into the general coding framework to ensure a comprehensive analysis. Notably, the findings from this coding framework revealed that participants found algorithmic systems to be mentally challenging,

often requiring significant cognitive effort to navigate and understand. Furthermore, the process of engaging with these systems heightened participants' awareness of the broader political conditions and implications associated with algorithmic decision-making. This increased awareness highlighted the intersection between technology and politics, illustrating how algorithmic decision-making can influence and be influenced by the political environment. These insights underscore the complex and multifaceted nature of algorithmic systems and their impact on individuals' cognitive and political landscapes. There is not enough data to make such an assumption, or it may be beyond the scope of this study. However, the increase in political awareness could potentially influence the emergence of new political movements in the domestic politics of countries in the long term. Similar to environmental parties, animal advocacy parties, political movements focusing on algorithmic systems and developing policies around them could emerge.

6. Conclusion and Future Research Directions

6.1.Introduction

This chapter was created to provide a general evaluation of our thesis, summarize the answer to the research question, and offer suggestions for future research. First, we will present the answer to our research question and then examine how justice can be addressed in algorithmic decision-making processes with Rawls' theory of justice. In this context, we will discuss how the concepts of justice and equity can be redefined or developed based on the results of the literature review. We will also consider how justice and equity can be reconceptualized in controversial uses of algorithmic decision-making processes. In light of these discussions, we will summarize our research question in a few sentences by offering a general answer.

Subsequently, after identifying the limitations of our study, possible research areas for future research will be shared. At the end of the chapter, I will share what we learned from this study and my advice for future Politics and Governance in the Digital Age Master's Program students (PGDA) students.

6.2.The Answer to the Research Question

According to the findings of the research, the three cases examined offer similar perspectives on the internal dynamics of the events. These results generally emphasize procedural and distributive justice. Based on the findings of the study, stakeholders of the public debate, whether immediate victims or those who cover the issue in academic or media sources, emphasize the importance of fair processes and fair distribution of resources. This situation offers an important perspective on the functioning of social and institutional structures in general.

6.2.1. Rawls Theory of Justice

Rawls's theory of justice is shaped around certain main concepts and is based on two principles. Reconsidering these concepts and principles in the context of algorithmic decision-making systems will contribute to the examination of cases.

According to Rawls, the concept of "basic structure" forms the basis of justice. This concept describes the way the institutions that make up the political and social structure come together and determines how they assign fundamental rights and distribute the social division of labor. The configuration of the fundamental framework delineates the opportunities and circumstances prevailing within the societal framework. It governs the allocation of resources, the execution of

processes, and analogous considerations. Therefore, the basic structure of a society constitutes the basis of justice in that society.

Likewise, the conditions shaping the basic structure will be an important factor in establishing justice in the context that algorithms are employed. These shaping conditions might be technological infrastructures, AI regulations, institutional frameworks, or among others. Adjusting the basic structure to a level that minimizes the potential side effects or harms of the algorithms will also ensure that the outputs are more fair.

In an environment where algorithmic decision-making systems are widespread, the identity of the basic structure concept and the institutions it contains are also highly related to these algorithms. For example, how do algorithmic systems in welfare programs affect people's rights, income inequality, and social justice? How might these systems contribute to making the underlying structure more just or more unequal? The sustainability of injustice or justice within the system depends, in a sense, on the relationship of the algorithms with the basic structure. Or questions such as how can we prevent the risks that algorithms may be biased against disadvantaged groups come to light? In the cases examined in the study, it was observed that the algorithms put into effect were not fair in the context of political and social institutions, both in the context of the process and the distribution of the results.

The basic structure is also closely related to how algorithms are designed. In this respect, in the design of algorithms, it is necessary to preserve the original position, as Rawls puts it, behind the veil of ignorance. Just as people must make decisions from behind the veil of ignorance to ensure the original position, those who design the algorithmic decision-making system must also preserve the original position. Although this is what should happen, it is almost impossible in the context of algorithm based decision making systems. Since the data themselves already contain bias, the original position is actually contaminated as soon as it starts. They are deprived of the ability to not know the characteristics that may cause prejudice. From this point of view, it becomes nearly impossible to determine base principles for justice. In this respect, it is almost impossible to go to behind the veil of ignorance and create a fair algorithm and ensure original position. Those who design or implement the system must take this into consideration. It can be clearly seen that this principle was violated in all three cases covered in the thesis.

Rawls, as mentioned in the 3th chapter, takes a position in relation to social contract theory. Thus, in a world where algorithms prevail may arise the question of how algorithms might affect or prompt a reevaluation of social contracts. With regard to social contract theory, it is suggested that there exist a fundamental agreement within society to establish a governance structure. Then, it is crucial to understanf what role algorithms play in establishing this structure. Without any doubt, they have a huge potential to alter political sphere and our understanding of public governance. In algorithmic decision-making processes, it is important that the decisions comply with the generally accepted values of the society. Although it is not possible for technical reasons to either maintain the original position or get behind the veil of ignorance in the context of designing and implementing algorithms, it is possible to observe the social contract. It is not possible to say that any social contract was observed in the cases examined in the study. In the SyRI case, the fact that the algorithm was applied only in certain areas raises doubts about whether any social contract was observed. In social contract theory, it is discussed how to ensure justice and equality, but in the relevant cases, rather than considering justice or equality, either certain groups of people were acted with the instinct of winning a war against an enemy who was not likely to be defrauded (social welfare recipients) or without taking into account the weaker strata of the society (students studying in public schools) algorithm has been developed.

With regard to two principles of Rawls, it is possible to say that algorithmic decision-making systems do not directly contradict the first principle; because no algorithm is designed to directly undermine these fundamental freedoms. However, if the algorithmic system conflicts with the second principle, fundamental freedoms may be eroded. The second principle generally deals with social and economic inequalities. The first article of this principle suggests providing accessible offices and positions to everyone (fair equality of opportunity), while the second article suggests regulating social inequalities for the benefit of the least advantaged.

The second principle is central to ongoing debates as it primarily addresses issues surrounding social and economic inequalities. While the first clause of this principle, which suggests that offices and positions should be accessible to everyone (fair equality of opportunity), is concerned with the distribution of opportunities, the second clause, which suggests that the inequalities of society should be arranged in a way that least advantaged strata are to be benefited, is concerned with procedural justice.

In the cases examined in this study, such as the A Level grading scandal, the existence of algorithms that prevent a certain group of people from accessing education and career opportunities is evident. Similarly, in cases such as Robodebt and SyRI, it has been observed that instead of adjusting social and economic inequalities in favor of the most disadvantaged group, the systems further deepen these inequalities and marginalize the weakest link of society. These cases show how easily algorithmic decision-making systems can reinforce existing injustices and disadvantage certain groups of society, rather than enhancing social justice.

In algorithmic decision making systems the expectation is mainly shaped around to be convenient and efficient. This may cause to overlook the concepts and principles put forward by Rawls. Algorithmic decision making systems may easily be deviated from what it ought to be, and given that these algorithmic arrangements may cause social unrest it is crucial to ensure the justice of both decisions and how they are distributed.

Especially when considered in the context of welfare states, the digitalization process negatively affects human rights. In this process, new risks and challenges arise in terms of digitalizing critical services such as social assistance programs, protecting human rights and ensuring social justice. With the use of algorithms, important principles such as transparency and accountability in decision-making processes can often be violated, and this can harm human rights. For example, algorithms making incorrect or biased decisions in the distribution processes in social assistance programs may lead to further marginalization of disadvantaged groups and violation of their rights. In this manner, UN Special Rapporteur on extreme poverty and human rights olan Philip Alston (2019) states that "*welfare state is gradually disappearing behind a webpage and an algorithm, with significant implications for those living in poverty*" (U.N. Human Rights Council, 2019: 13). Further, he adds that '*in the absence of transparency about the existence and workings of automated systems, the rights to contest an adverse decision and to seek a meaningful remedy are illusory*' (U.N. Human Rights Council, 2019, p. 15).

6.2.2. Fairness in Algorithmic Decision-Making Systems

The literature review demonstrates that the implementation of artificial intelligence in government is examined in terms of fairness, considering distributive justice and procedural justice perspectives. Procedural justice focuses on whether the decision-making process is fair, encompassing principles such as transparency, accountability, and impartiality. Distributive justice,

on the other hand, scrutinizes whether the outcomes are fair and aims to ensure the equal distribution of resources or benefits among individuals.

Based on this, the reconceptualization and refinement of the fairness concept can be achieved within a context where the process of implementation and the concept of distributive justice are not considered separately. This approach not only entails designing algorithms that produce fair outcomes but also ensuring they adhere to fair decision-making processes. Moreover, considering contextual factors such as historical biases, societal factors, and structural inequalities can further enhance the fairness of algorithmic systems.

In the context of algorithmic decision-making, justice can also be reconceptualized by highlighting administrative and legal regulations in addition to, or even complementing, the aspects mentioned above. Algorithms continue to impact various aspects of our lives, presenting us with a new philosophy in a sense. The opportunities or conditions brought forth by this philosophy should be thoroughly examined from a legal standpoint.

6.2.3. Overall Answer

The concept of fairness surfaced from data in a similar way in all three cases. The sense of justice in public debates reflects similar concerns. Worries about lack of transparency and accountability are evident in all three cases. Privacy violations are also present in both the Robodebt and SyRI cases, which are social welfare fraud detection systems. Additionally, issues like bias and discrimination are prominent in discussions about fairness in the data.

Based on this, we can conclude that distributive and procedural justices are prominent concerns in public debates about algorithmic decision-making scandals. Although public debates took place over the results of a specific algorithmic decision making systems, the processes that lead to those results also find their place in those discussions. Algorithms alone did not worsen the existing situation; they only made the problem more visible and accessible. So to speak, they turned the glass ceiling into a wall that everyone could see. This requires fairness to be redefined by algorithmic systems. Redefining fairness will prevent defining algorithmic systems solely around efficiency from inviting unexpected problems. This requires ensuring transparency, accountability and due process in algorithmic decision-making processes, while striving for equitable outcomes that promote social justice and protect fundamental rights. Ethical issues such as protection of privacy, non-discrimination and respect for human dignity should also be taken into account when

designing algorithmic systems. This necessitates a definition of justice that includes not only technical aspects but also ethical dimensions that prioritize human rights and social values.

6.3.Limitations

This study has some limitations that will not affect its overall reliability.

First of all, the study was restricted to documents and sources in English, significantly limiting the available documentation on the Dutch SyRI case. Given that the study employs thematic analysis of textual data, this linguistic difference might affect the analysis of this case. It is likely that debates conducted in Dutch contain nuances that could enrich the study's findings.

The study examines scandals related to algorithmic decision-making systems. By studying scandals, impacts are easier to observe. This is a valid reason to design the study around scandals. However, the fairness issues are likely overstated under such a condition and should not be generalized to other AI cases. This might be the case for small-n studies; in this study, one should be aware that cases are composed of scandals.

Similarly, the data for the Robo-debt case provides a direct reflection of the victims' feelings and perceptions of the issue, particularly through their e-petitions. These petitions served as a means for individuals to voice their concerns and push both the public and the Royal Commission in their struggle against the program. While it is not suggested that the petitioners were dishonest, there is a possibility that they might have exaggerated their emotions to make a contribution to the developments surrounding the Royal Commission. Nevertheless, the e-petitions were the most effective tool available to understand the implications of Robo-debt.

Lastly, although cases were chosen based on specific criteria, the time frames of the studies do not fully overlap, which may affect the comparability of the cases.

6.4.Future research directions

After addressing the research question, it's evident how this study contributed to an ongoing discourse in academic literature regarding fairness in relation to algorithms. As with any research endeavor, this study has its limitations, as outlined in the preceding section. However, while conducting the study, and particularly in developing the theoretical framework and analyzing empirical data, new inquiries and deficiencies emerged. In the following discussion, the most significant ones are addressed and connected to potential avenues for future research.

Mental health-related problems might indicate a broader research direction in the emerging field of human and non-human interactions. The increasing use of artificial intelligence directly or indirectly affects mental health. Artificial intelligence has been associated with potential risks to mental health, indirectly leading to conditions such as loneliness, insomnia, and lack of self-confidence. Even in cases where it does not directly affect mental health, it can trigger side effects that can negatively affect mental health, as seen in this study. The Royal Australian and New Zealand College of Psychiatrists (RANZCP, 2023), in its report, Improve the Mental Health of Communities, strongly recommends that individuals affected by the program receive complimentary mental health care that is culturally sensitive, aiming to alleviate any trauma-related effects.

I propose considering these mental health issues as a future research direction not merely as problems faced by individuals experiencing economic hardships, but rather as societal phenomena akin to Blauner's concept of alienation in relation to technology. Blauner's concept of alienation has generally been used in work organizations but can be applied to society as well. This concept is used to explain the increasing alienation of workers' relationships with their jobs, especially in industrial enterprises, due to technological developments in the production process and division of labor.

Blauner's concept of alienation, stemming from the shifts brought about by industrialization and urbanization in modern society (Kanungo, 1979), offers insights into potential parallels with the challenges posed by algorithmic decision-making systems. This is primarily due to the fact that individuals subjected to alienation during Blauner's era often experienced a detachment from society and their surroundings, marked by feelings of disconnection and insecurity. Similarly, individuals engaging with algorithmic systems today may encounter similar sensations of alienation.

Given the rapid pace of technological advancement in contemporary society, many individuals may lead isolated lives with minimal interpersonal interaction. Furthermore, when confronted with artificial intelligence and a sense of powerlessness in navigating its complexities, individuals may experience heightened alienation from their community, feelings of vulnerability, and a diminished sense of belonging.

Alienation has been a topic of extensive research across various disciplines. However, contemporary advancements in artificial intelligence offer fresh avenues for exploring this phenomenon from novel vantage points.

The possibility that such symptoms indicate a sociological tendency rather than being interpreted as experiences specific to individuals should not be ignored. Alienation is influenced not only by individual psychological factors but also by social structures and dynamics. In this context, I believe that the mental health problems identified in the study are also societal issues and deserve to be examined within the framework of algorithms and artificial intelligence.

Exploring Rawls's concept of the original position in the context of algorithmic decision-making systems is worth considering. The original position assumes a scenario where all actors are equal, with none having advantages or disadvantages over others. One of the critical elements of the original position is the veil of ignorance, which ensures that individuals are unaware of their personal characteristics, social status, wealth, abilities, gender, race, and other personal conditions. This allows for impartial and objective decision-making regarding just societal principles. Is it possible to achieve this condition in data-driven, algorithmic decision-making systems? Or can the veil of ignorance and the original position be realized in algorithmic decision-making systems?

6.5.Epilogue

As I conclude this master's thesis, I would like to share what I have learned throughout this journey and what advice I would like to offer to the students of the next generation of PGDA master's program. The research conducted, the many articles read, the intense discussions in seminars and lectures, and the moments of doubt and enlightenment have deeply influenced this work and my understanding of the subject. During this process, I encountered many challenges, from conceptualizing research questions to analyzing data. However, each challenge also presented a learning opportunity and encouraged me to think more analytically and solution oriented. The achievements, although sometimes small, were extremely rewarding and strengthened my commitment to this research.

One of the most important things I learned during this process is that research is an endeavor that requires patience and discipline. Each stage of research requires careful planning and attention to detail.

In consideration of these points, I would like to proceed with offering advice to future PDGA students. Although the use of data analysis tools is becoming more user-friendly and easier day by day, data analysis is still a process that involves many details and requires quality researchers. Having students specialize in both qualitative and quantitative data analysis as much as possible will facilitate their thesis writing process and contribute to their researcher identities.

Estonia is one of the best countries in the world to study digital governance and offers many event opportunities in this context. Following these studies and participating, if possible, will contribute to their knowledge of the field. These events also provide network development opportunities.

Another recommendation I have for PDGA students is to form a group with friends who share common interests and create a platform where they can share developments, news, or academic studies they find interesting with each other. They can convene regularly to engage in discussions. These gatherings, irrespective of the formal demands of the master's program, will foster a comprehensive understanding of the subject matter.

Interdisciplinary collaboration and integrating perspectives from various disciplines will broaden the understanding of digital governance, allowing for a more comprehensive analysis of its multifaceted challenges and opportunities. By drawing upon insights from fields such as political science, computer science, public administration, sociology, and ethics, students can gain a deeper appreciation of the complex interplay between technological innovation, policy development, social dynamics, and ethical considerations in the digital age. This holistic approach not only enriches academic inquiry but also equips future leaders and policymakers with the interdisciplinary skills and perspectives necessary to address the evolving complexities of governance in an increasingly digitalized world.

Ultimately, I highly advocate for prospective PGDA students to explore internship or employment prospects within the domain of digital governance. Such experiences afford them the chance to witness the practical application of their acquired knowledge from study groups and academic pursuits. This not only enhances their postgraduate academic or professional trajectories but also enriches their educational journey, fostering a more fulfilling learning experience.

References

Alhojailan, M. I. (2012). Thematic analysis: a critical review ofits process and evaluation. In WEI international European academic conference proceedings, Zagreb, Croatia.

AlgorithmWatch. (2020). How Dutch activists got an invasive fraud detection algorithm banned. AlgorithmWatch. https://algorithmwatch.org/en/syri-netherlands-algorithm/

Allegretti, A. (2020, August 13). A-level results: Government accused of 'baking in' inequality with 'boost' for private schools. Sky News. https://news.sky.com/story/35-of-a-level-results-downgraded-by-one-grade-figures-reveal-12048251

Barsotti, F., & Koçer, R. G. (2022). MinMax fairness: from Rawlsian Theory of Justice to solution for algorithmic bias. *AI & SOCIETY*, 1-14.

Bedingfield, W. (2023, February 10). How the A-Level Exam Algorithm Screwed Students. Wired. https://www.wired.co.uk/article/alevel-exam-algorithm

BBC. (2020, August 19). Durham University students offered money to defer after exams U-turn. BBC News. https://www.bbc.com/news/uk-england-tyne-53840379

BBC. (2020, August 20). A-levels and GCSEs: How did the exam algorithm work? BBC News. https://www.bbc.com/news/explainers-53807730

BBC News. (2020, August 25). Ofqual chief Sally Collier steps down after exams chaos. Retrieved from https://www.bbc.com/news/education-53909487

Broomfield, H., & Reutter, L. M. (2021). Towards a data-driven public administration: An empirical analysis of nascent phase implementation.

Carney, T. (2018, June 5). Social security and Robo-debt. Austaxpolicy. https://www.austaxpolicy.com/social-security-robo-debt/

Chong, Y. W., Villanueva-Libunao, K., Chee, S. Y., Alvarez, M. J., Yau, K. L. A., & Keoh, S. L. (2022). Artificial intelligence policies to enhance urban mobility in Southeast Asia. *Frontiers in Sustainable Cities*, *4*, 824391.

Coeckelbergh, M. (2022). The Political Philosophy of AI. Polity Press.

Colclough, C. J., & Lappin, K. (2023). Building Union Power to Rein in the AI Boss. Stanford Social Innovation Review. https://doi.org/10.48558/995G-CJ51

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 797-806).

Danaher, J. (2022). Techno-optimism: An analysis, an evaluation and a modest defence. Philosophy & Technology, 35(2), 54.

Drake, A., Keller, P., Pietropaoli, I., Puri, A., Maniatis, S., Tomlinson, J., ... & Blair, S. W. (2022). Legal contestation of artificial intelligence-related decision-making in the United Kingdom: Reflections for policy. *International Review of Law, Computers & Technology*, *36*(2), 251-285.

Donald Moon, J. (2015). Cohen vs. Rawls on justice and equality. *Critical Review of International Social and Political Philosophy*, *18*(1), 40-56.

Dyer-Witheford, N., Kjøsen, A. M., & Steinhoff, J. (2019). Inhuman Power: Artificial Intelligence and the Future of Capitalism. Pluto Press.

European Commission. (2018). Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions: Artificial Intelligence for Europe.

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. Journal of advanced nursing, 62(1), 107-115.

Erman, E., & Furendal, M. (2022). The global governance of artificial intelligence: some normative concerns. *Moral Philosophy and Politics*, *9*(2), 267-291.

Faraj S., Pachidi S., and Sayegh, K. Working and organizing in the age of the learning algorithm. (2018). Information and Organization, 28(1), 62–70. Available at:. https://doi. org/10.1016/j.infoandorg.2018.02.005

Farnadi, G., Babaki, B., & Getoor, L. (2018, December). Fairness in relational domains. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 108-114)

Ferretti, T. (2022). An institutionalist approach to AI ethics: justifying the priority of government regulation over self-regulation. *Moral Philosophy and Politics*, *9*(2), 239-265.

Follesdal, A. (2014). John Rawls' theory of justice as fairness. In *Philosophy of Justice* (pp. 311-328). Dordrecht: Springer Netherlands.

Fuchs, C. (2014). Social Media: A Critical Introduction. SAGE Publications Ltd.

Human Rights Watch. (2019, November 8). Welfare surveillance on trial in the Netherlands. Retrieved from https://www.hrw.org/news/2019/11/08/welfare-surveillance-trial-netherlands

Heikkilä, M. (2022, March 29). Dutch scandal serves as a warning for Europe over risks of using algorithms. Politico Europe. Retrieved from https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/

Hinton, C. (2023). The State of Ethical AI in Practice: A Multiple Case Study of Estonian Public Service Organizations. *International Journal of Technoethics (IJT)*, *14*(1), 1-15.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. Qualitative health research, 15(9), 1277-1288.

Ienca, M. (2019). Democratizing cognitive technology: a proactive approach. Ethics and Information Technology, 21(4), 267-280.

Jiang, Y., Li, X., Luo, H., Yin, S., & Kaynak, O. (2022). Quo vadis artificial intelligence?. *Discover Artificial Intelligence*, *2*(1), 4.

Johnson, S., & Acemoglu, D. (2023). *Power and progress: Our thousand-year struggle over technology and prosperity*. Hachette UK.

Jones, C. (2023, January 31). Robodebt victims' details released to media in bid to deter them from speaking out, royal commission told. ABC News. Retrieved from https://www.abc.net.au/news/2023-01-31/qld-robodebt-scheme-government-royal-commission-alan-tudge/101910064

Kanungo, R. N. (1979). The concepts of alienation and involvement revisited. *Psychological bulletin*, *86*(1), 119.

Karp, P. (2023, July 9). Early robodebt critics outraged by how long Coalition persisted with unlawful scheme. The Guardian. Retrieved from https://www.theguardian.com/australia-

news/2023/jul/09/early-robodebt-critics-outraged-by-how-long-coalition-persisted-with-unlawful-scheme

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, *13*(3), 795-848.

Kuziemski, M., & Misuraca, G. (2020). AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. Telecommunications policy, 44(6), 101976.

Leventhal, G. S. (1980). What should be done with equity theory? New approaches to the study of fairness in social relationships. In *Social exchange: Advances in theory and research* (pp. 27-55). Boston, MA: Springer US.

Le Sueur, A. (2015). Robot government: automated decision-making and its implications for parliament. *This is a draft chapter, to be published in A Horne and A Le Sueur (ed), Parliament: Legislation and Accountability*.

Luetjens, J., Mintrom, M., & t Hart, P. (2019). Successful public policy: Lessons from Australia and New Zealand. ANU Press.

Lovett, F. (2010). *Rawls's' A theory of justice': A reader's guide*. Bloomsbury Publishing.

Marks, R. (2020, November 18). Robodebt: How did it come to this? The Monthly. Retrieved from https://www.themonthly.com.au/blog/russell-marks/2020/18/2020/1605669075/robodebt-how-did-it-come

Mead, G., & Barbosa Neves, B. (2023). Contested delegation: Understanding critical public responses to algorithmic decision-making in the UK and Australia. *The Sociological Review*, *71*(3), 601-623.

Newell, S., & Marabelli, M. (2020). Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. In Strategic Information Management (pp. 430-449). Routledge.

LexisNexis. (2024). About Us. Retrieved May 1, 2024, from https://www.lexisnexis.com/en-us/about-us/about-us.page

Namey, E., Guest, G., Thairu, L., & Johnson, L. (2008). Data reduction techniques for large qualitative data sets. *Handbook for team-based qualitative research*, *2*(1), 137-161.

OECD. (2019). Recommendation of the Council on OECD Legal Instruments Artificial Intelligence. Paris. Retrieved from https://www.oecd.ai/ai-principles

Ofqual. (2020, September 2). Written statement from Chair of Ofqual to the Education Select Committee. Ofqual. Retrieved from https://www.gov.uk/government/news/written-statement-from-chair-of-ofqual-to-the-education-select-committee

O'Neil, C. (2016). Weapons of Math Destruction. Crown Publishers.

Orji, C. P. (2023). THE SOCIAL CONTRACTS FROM THE LENS OF JOHN RAWLS VEIL OF IGNORANCE: A PHILOSOPHICAL ANALYSIS. *IGWEBUIKE: African Journal of Arts and Humanities*, *9*(4).

Parker, K. (2020, March 20). GCSE resits: students to receive 'calculated' grades. Tes Magazine. https://www.tes.com/magazine/archive/gcse-resits-students-receive-calculated-grade

Parliament of Australia. (2017, June 21). Design, scope, cost-benefit analysis, contracts awarded and implementation associated with the Better Management of the Social Welfare System initiative.

Polevikov, S. (2023). Advancing AI in healthcare: a comprehensive review of best practices. Clinica Chimica Acta, 117519.

PRISMA. (n.d.). PRISMA statement. Retrieved from http://www.prisma-statement.org/

Rawls, J. (2001). *Justice as fairness: A restatement*. Harvard University Press.

Rachovitsa, A., & Johann, N. (2022). The human rights implications of the use of AI in the digital welfare state: Lessons learned from the Dutch SyRI case. *Human Rights Law Review*, *22*(2), ngac010.

Rafanelli, L. M. (2022). Justice, injustice, and artificial intelligence: Lessons from political theory and philosophy. *Big Data & Society*, *9*(1), 20539517221080676.

Riger, S. and Sigurvinsdottir, R. (2016) Thematic Analysis, in Jason, L.A. and Glenwick, D.S. (eds.) Handbook of Methodological Approaches to Community-Based Research: Qualitative, Quantitative, and Mixed Methods. New York, NY: Oxford University Press, pp. 33-41.

Rinta-Kahila, T., Someh, I., Gillespie, N., Indulska, M., & Gregor, S. (2023). Managing unintended consequences of algorithmic decision-making: The case of Robodebt. Journal of Information Technology Teaching Cases, 20438869231165538.

Royal Australian and New Zealand College of Psychiatrists. (2023). Improve the mental health of communities: Submission to the Royal Commission into the Robodebt Scheme.

Schreier, M. (2012). Qualitative Content Analysis in Practice. SAGE Publications Ltd.

Schiff, D. S., Schiff, K. J., & Pierson, P. (2022). Assessing public value failure in government adoption of artificial intelligence. *Public Administration*, *100*(3), 653-673.

Seabright, P., Stieglitz, J., & Van der Straeten, K. (2021). Evaluating social contract theory in the light of evolutionary social science. *Evolutionary Human Sciences*, *3*, e20.

Sharifi-Malvajerdi, S., Kearns, M., & Roth, A. (2019). Average individual fairness: Algorithms, generalization and experiments. *Advances in neural information processing systems*, *32*.

Shead, S. (2020, August 21). Computer algorithm caused a grading crisis in British schools. CNBC. Retrieved from https://www.cnbc.com/2020/08/21/computer-algorithm-caused-a-grading-crisis-in-british-schools.html

Sky News. (2020, August 16). A-levels: How controversial algorithm behind moderation row works. Sky News. https://news.sky.com/story/a-levels-how-controversial-algorithm-behind-moderation-row-works-12048780

Stein, S. M., & Harper, T. L. (2005). Rawls's 'justice as fairness': A moral basis for contemporary planning theory. *Planning Theory*, *4*(2), 147-172.

Strauss, A., & Corbin, J. (1998). Basics of qualitative research techniques.

Submission to the Royal Commission into Robodebt. (n.d.). Retrieved February 7, 2024, from https://robodebt.royalcommission.gov.au/submission

Torbert, W. (1974). Doing Rawls justice. *Harvard Educational Review*, *44*(4), 459-469.

Turque, B. (2012). Creative ... motivating. The Washington Post. Retrieved from [https://www.washingtonpost.com/local/education/creative--motivating-and fired/2012/02/04/gIQAwzZpvR_story.html]

UK Parliament. (2020, February 12). Algorithms in Public Sector Decision-Making. Hansard. https://hansard.parliament.uk/lords/2020-02-12/debates/C8998608-FA12-4BFC-BEAE-4057D699FDBD/AlgorithmsPublicSectorDecision-Making

U.N. Human Rights Council (2019) Visit to the United Kingdom of Great Britain and Northern Ireland. A/HRC/41/39/Add.1, Report of the Special Rapporteur on extreme poverty and human rights, 23 April. Available at: https://ap.ohchr.org/documents/dpage_e.aspx?si¼A/HRC/41/39/Add.1.

Utting, A. (2023, March 7). Top bureaucrat who oversaw Robodebt rollout says document to ministers was potentially misleading. ABC News. Retrieved from https://www.abc.net.au/news/2023-03-07/robodebt-cabinet-document-misleading-kathryn-campbell/102064104

Van Bekkum, M., & Borgesius, F. Z. (2021). Digital welfare fraud detection and the Dutch SyRI judgment. European Journal of Social Security, 23(4), 323-340.

Xivuri, K., & Twinomurinzi, H. (2021). A systematic review of fairness in artificial intelligence algorithms. In Responsible AI and Analytics for an Ethical and Inclusive Digitized Society: 20th IFIP WG 6.11 Conference on e-Business, e-Services and e-Society, I3E 2021, Galway, Ireland, September 1–3, 2021, Proceedings 20 (pp. 271-284). Springer International Publishing.

Wang, X., Li, L., Tan, S. C., Yang, L., & Lei, J. (2023). Preparing for AI-enhanced education: Conceptualizing and empirically examining teachers' AI readiness. *Computers in Human Behavior*, *146*, 107798.

Wang, R., Harper, F. M., & Zhu, H. (2020, April). Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-14).

Whyte, S. (2019, February 18). Minister denies robodebt caused more than 2000 deaths. The Sydney Morning Herald. Retrieved from https://www.smh.com.au/politics/federal/minister-denies-robodebt-caused-more-than-2000-deaths-20190217-p50yep.html

Wieringa, M. (2023). "Hey SyRI, tell me about algorithmic accountability": Lessons from a landmark case. *Data & Policy*, *5*, e2.

Zimmerman, J. (2022). Small N. In J. Clipperton et al. (Eds.), Empirical Methods in Political Science: An Introduction. Northwestern University Libraries. Retrieved from https://nulib-oer.github.io/empirical-methods-polisci/small-n.html

Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, *31*(4), 1060-1089.