CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Sociological Studies
Department of Public and Social Policy

Master's Thesis

2025 Tigran Karapetyan

CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Sociological Studies
Department of Public and Social Policy

Accountability in Algorithmic Decision Making: Case Study Analysis

Master's Thesis

Author of the Thesis: Tigran Karapetyan

Study programme: Public and Social Policy

Supervisor: Mirna Jusić, M.A., Ph.D.

Year of the defence: 2025

Declaration

1. **I hereby declare that I have compiled this thesis using the listed literature and resources only.**
2. **I hereby declare that my thesis has not been used to gain any other academic title.**
3. **I fully agree to my work being used for study and scientific purposes.**
4. **During the preparation of this thesis, the author used the AI tool *ChatGPT* developed by OpenAI in order to support text formulation, structure clarification, and reference formatting. All outputs generated by the tool were critically reviewed, edited, and verified by the author, who takes full responsibility for the accuracy and academic integrity of the final content.**

In Prague on **Tigran Karapetyan**
28.07.2025

References

KARAPETYAN, Tigran. Accountability in Algorithmic Decision Making: Case Study Analysis. Praha, 2025. 63 p. Master's thesis (Mgr). Charles University, Faculty of Social Sciences, Institute of Sociological Studies, Department of Public and Social Policy. Supervisor Mirna Jusić, M.A., Ph.D.

Length of the Thesis: 129,797 Characters

Abstract

This thesis investigates the accountability of algorithmic decision-making (ADM) systems in public sector governance, using the Dutch childcare benefits scandal as a case study. The research examines how an ADM system implemented by the Dutch Tax Authority wrongfully flagged thousands of families as fraudulent, leading to severe social and financial harm. Through a qualitative methodology combining interviews and secondary data, the study applies Nicholas Diakopoulos's Algorithmic Accountability Framework, focusing on fairness, transparency, and explainability. The findings reveal systemic failures across all three dimensions, including discriminatory data proxies, lack of procedural clarity, and institutional opacity. Furthermore, the analysis highlights the broader socio-political dynamics that enabled these accountability gaps, drawing on the concept of algorithmic governance. The study concludes that technical tools alone cannot ensure accountability; rather, strong institutional oversight, inclusive design processes, and well-defined procedures for transparency and redress are essential. Lessons from this case offer insights for future ADM implementations in the public sector, emphasizing the urgent need for frameworks that safeguard citizens' rights and promote just governance.

Keywords

Algorithmic Decision-Making, ADM, Algorithmic Accountability, Dutch Childcare, Fairness, Transparency, Explainability, Institutional Oversight, algorithmic governance

Abstrakt

Tato diplomová práce zkoumá odpovědnost systémů algoritmického rozhodování (ADM) ve veřejné správě na příkladu skandálu s dětskými přídavky v Nizozemsku. Výzkum analyzuje, jak systém ADM, zavedený nizozemským daňovým úřadem, neoprávněně označil tisíce rodin za podvodné žadatele o dávky, což vedlo k závažným sociálním a finančním dopadům. Pomocí kvalitativní metodologie, která kombinuje rozhovory a sekundární data, práce aplikuje Rámec algoritmické odpovědnosti Nicholase Diakopoulose se zaměřením na spravedlnost, transparentnost a srozumitelnost. Zjištění odhalují systémová selhání ve všech třech oblastech, včetně diskriminačních datových proxy proměnných, nedostatečné procesní jasnosti a institucionální neprůhlednosti. Analýza zároveň poukazuje na širší sociálně-politické dynamiky, které tyto nedostatky umožnily, a opírá se o koncept algoritmického vládnutí. Studie dochází k závěru, že samotné technické nástroje k zajištění odpovědnosti nestačí; nezbytný je silný institucionální dohled, inkluzivní návrh systémů a jasně definované postupy pro zajištění transparentnosti a nápravy. Případ poskytuje cenné poznatky pro budoucí využívání ADM ve veřejném sektoru a zdůrazňuje naléhavou potřebu rámců, které chrání práva občanů a podporují spravedlivé vládnutí.

Klíčová slova

Algoritmické rozhodování, ADM, algoritmická odpovědnost, nizozemský případ dětských přídavků, spravedlnost, transparentnost, srozumitelnost, institucionální dohled, algoritmické vládnutí

Název práce

Odpovědnost v algoritmickém rozhodování: analýza případové studie

Table of Content

This page was left blank intentionally.

1. Introduction

Artificial intelligence (AI) and machine learning (ML) have become a central topic in discussions about technological advancements and societal change. According to the Oxford dictionary (Oxford University Press, 2025), AI is “the study and development of computer systems that can copy intelligent human behavior” and ML is “a type of artificial intelligence in which computers use huge amounts of data to learn how to do tasks rather than being programmed to do them”. AI consists of algorithms trained on data that enable a computer to make decisions based on predefined criteria. With machine learning, AI systems can improve over time through feedback loops, leading to increasingly precise decision-making. Businesses have been using AI for years to enhance efficiency. It is often assumed that AI-driven decision-making is more objective and fairer, as machines should, in theory, be free from human biases (O’Neil, 2016; Eubanks, 2018; Barocas, Hardt, & Narayanan, 2019). However, is this truly the case?

In the private sector, AI is widely used for tasks such as evaluating loan applications in banking or screening job applications in recruitment. AI-driven credit scoring models, such as those used by JPMorgan Chase and Goldman Sachs, assess loan applications based on financial history, spending patterns, and alternative data sources (Fuster et al., 2022). Companies like HireVue and Pymetrics use AI to analyze candidates’ resumes, video interviews, and behavioral traits to match them with job requirements (Raghavan et al., 2020). Instead of human employees manually reviewing thousands of requests, AI systems assess applications based on programmed criteria, selecting the most eligible candidates. This automation increases efficiency and consistency in decision-making.

The public sector is also increasingly adopting AI to streamline processes (AlgorithmWatch, 2020; OECD, 2021). The Dutch childcare benefits scandal (2013–2020) is a notable failure, where an AI-driven fraud detection system falsely accused thousands of families, leading to severe financial and social consequences (Veale & Brass, 2019). The Los Angeles Police Department (LAPD) used PredPol, an AI tool predicting crime locations based on historical data. Studies found it disproportionately targeted minority communities, reinforcing existing biases rather than reducing crime (Richardson et al., 2019). In the UK, Durham Police tested the Harm Assessment Risk Tool (HART) to assess the likelihood of reoffending (Oswald et al., 2018). Unlike private businesses, which often have more flexibility in designing their AI systems without public accountability, government agencies must ensure transparency, fairness, and responsibility in their decision-making. When AI is used to influence or determine public decisions, ethical and democratic concerns arise, requiring robust oversight and accountability structures (Veale & Brass, 2019; Yeung, Howes, & Pogrebna, 2019; Ziewitz, 2016). As Veale and Brass (2019) argue, algorithmic decision making (ADM) in the public sector challenges traditional governance models and raises questions about delegated authority and bureaucratic legitimacy. Similarly, Yeung et al. (2019) warn that if algorithmic systems unchecked, they can obscure accountability and reduce opportunities for public contestation.

Accountability in government refers to the obligation of public officials and institutions to be answerable for their actions, decisions, and policies to the public and oversight bodies. It ensures transparency, responsiveness, and adherence to legal and ethical standards, preventing corruption and misuse of power (Bovens, 2007). Effective accountability in government increases trust among people towards government, improves policy outcomes, and upholds the principles of democracy (Dubnick, 2005). In the context of algorithmic decision-making, accountability becomes even more critical, referring to ensuring that AI-driven systems and their outcomes remain transparent, explainable, and subject to oversight. Since ADM often involves automated or semi-automated processes that impact individuals' lives, clear mechanisms must be in place (Veale & Brass, 2019). Governments must clarify how and why AI-driven decisions are made and establish mechanisms to assign responsibility for errors or biases. Without proper oversight, ADM systems can be manipulated, misused for private gain, or generate harmful unintended consequences.

Multiple studies and governmental reports indicate that the use of automated decision-making tools and artificial intelligence systems in the public sector has been steadily growing (AlgorithmWatch, 2020; OECD, 2021; European Commission, 2019; Amnesty International, 2021). The OECD has published reports on AI and digital government suggesting that a growing number of member countries are integrating

ADM into public services (OECD, 2019; OECD, 2021). The European Commission's eGovernment Benchmark reports an increasing adoption of digital tools, including automated systems and AI solutions, in the delivery of public services (European Commission, 2020). AlgorithmWatch, a European NGO, publishes the Automating Society reports (e.g., 2020, 2021 editions) mapping the use of ADM systems in various European countries. They document a clear upward trend in adoption, ranging from welfare fraud detection to resource allocation in healthcare.

Unlike traditional administrative processes, algorithmic decision-making often operates within opaque "black boxes" that obscure not only how decisions are made but also who is responsible for them (Pasquale, 2015). A black box refers to a system or process whose internal workings are not visible or understandable to the user, even though its inputs and outputs are known (Pasquale, 2015). This opacity complicates the basic democratic principle of answerability: if no clear human actor can be held accountable for an unjust decision, then meaningful oversight becomes elusive. In public governance, where decisions directly impact people's lives, the lack of transparency undermines both institutional trust and procedural legitimacy (Diakopoulos, 2015; Kroll et al., 2017).

The moral and democratic stakes of ADM are particularly pronounced in the public sector, where citizens cannot "opt out" of government systems in the way they might with private services. As Diakopoulos (2016) and Barocas et al. (2019) argue, the deployment of ADM technologies without adequate safeguards can erode essential values such as equal treatment and non-discrimination. Moreover, ADM challenges the foundational accountability mechanisms of public administration by introducing systems whose decision logic may be unintelligible not only to the public but even to those implementing them (Veale & Brass, 2019; Kroll et al., 2017). Therefore, investigating the fairness and accountability of ADM in state decision-making is not only a matter of technical evaluation but also an inquiry into the future of democratic governance (Ziewitz, 2016; Yeung, 2018).

The rise of algorithmic decision-making has often been accompanied by optimistic claims about objectivity, efficiency, and the elimination of human bias (Pasquale, 2015). Algorithms, unlike people, do not suffer fatigue, emotional variability, or cognitive inconsistencies. Governments and organizations have thus embraced ADM with the belief that it can streamline bureaucratic procedures, reduce administrative errors, and produce fairer outcomes by relying on “neutral” data-driven models (O’Neil, 2016; Eubanks, 2018; Barocas, Hardt, & Narayanan, 2019). This narrative has been particularly appealing in areas where decision-making is high-volume and resource-constrained, such as welfare, immigration, and law enforcement (Benjamin, 2019; Ziewitz, 2016).

Yet this promise masks a fundamental pitfall: data are not neutral, and neither are the institutions that generate and use them. As Barocas, Hardt, and Narayanan (2019) emphasize, ADM systems trained on historical or institutional data often reproduce the very patterns of bias and inequality they are meant to overcome. In practice, seemingly neutral variables can act as proxies for protected characteristics, leading to disparate impacts that are difficult to detect and challenge. For example, an algorithm might exclude race as an input, but still use postal codes, which can strongly correlate with race due to historical segregation patterns—thus enabling indirect discrimination. These kinds of proxy variables allow systems to replicate bias even when sensitive information is not explicitly used (Barocas, Hardt, & Narayanan, 2019).

Moreover, ADM can create self-reinforcing feedback loops, where biased outputs inform future decision-making, further entrenching discrimination over time (Richardson, Schultz, & Crawford, 2019). These issues are particularly important in public sector applications, where decisions often carry life-changing consequences and where individuals affected by these systems frequently lack the tools, knowledge, or legal standing to understand or contest algorithmic outcomes (O’Neil, 2016; Pasquale, 2015).

Algorithmic decision-making in the public sector rarely occurs in a vacuum. In many cases, ADM systems are implemented as decision-support tools rather than fully autonomous systems. This means that public officials interact with, interpret, and often rely on algorithmic outputs to make final determinations (Kroll et al., 2017; Veale & Brass, 2019; Ziewitz, 2016). These hybrid arrangements introduce new cognitive and behavioral dynamics that can significantly shape outcomes. Research by Alon-Barkat and Busuioc (2023) highlights two critical patterns: automation bias, where human decision-makers over-trust algorithmic outputs even in the face of contradictory evidence, and selective adherence, where users tend to follow algorithmic advice when it confirms their existing stereotypes or expectations.

In the Dutch childcare benefits scandal, for instance, civil servants often acted on algorithmic risk scores without fully understanding or questioning their basis (Amnesty International, 2021; Parliamentary Interrogation Committee on Childcare Allowance, 2020). As a result, even semi-automated systems became de facto gatekeepers to vital welfare services. This human–algorithm entanglement complicates traditional notions of responsibility and challenges existing accountability frameworks (Diakopoulos, 2015). It also underscores that ensuring fairness in ADM requires addressing both technological design and the institutional cultures and incentives surrounding its use (Veale & Brass, 2019).

1.2 Historical development of ADM and the Dutch childcare benefits scandal

Algorithmic Decision-Making in public administration has evolved significantly over the last several decades, shifting from basic digital tools to highly complex algorithmic systems. Initially, from the 1960s to the 1980s, public administrations began integrating basic information technologies primarily for administrative tasks, such as payroll management, tax processing, and record-keeping. These early systems were mostly rule-based, supporting rather than replacing human decision-making processes (Snellen, 2005).

During the 1980s and 1990s, public institutions increasingly experimented with expert systems. These systems employed predefined rules to automate routine decisions, notably in areas such as tax auditing, welfare eligibility determinations, and other bureaucratic functions. Although relatively simple, these early automated systems set the foundation for more complex, data-driven decision-making tools (Lipsky, 1980).

Entering the 2000s, the rise of e-government initiatives accelerated the adoption of data-driven ADM. These initiatives aimed to enhance efficiency, transparency, and public service delivery. Governments utilized automated systems extensively for risk profiling, fraud detection, immigration control, social welfare programs, and taxation purposes. This shift represented a significant step toward relying more heavily on technology to directly inform or make critical administrative decisions (Margetts & Dunleavy, 2013).

The last decade, from 2010 onwards, marked a dramatic shift as advancements in machine learning and the proliferation of big data enabled governments to deploy sophisticated ADM systems on a much larger scale. These advanced systems appeared in various public sectors, including predictive policing, welfare eligibility checks, child protection services, and fraud detection. However, this expansion brought significant ethical concerns and controversies, especially regarding algorithmic biases, opacity, and challenges associated with ensuring transparency and accountability (Eubanks, 2018; Yeung, 2018).

The Dutch childcare benefits scandal represents a notable example of the risks inherent in automated decision-making within public administration. Between approximately 2013 and 2019, automated systems implemented by the Dutch Tax Authority wrongly accused thousands of families, many of whom had minority backgrounds, of committing fraud related to childcare allowances. These allowances are financial subsidies provided by the government to help parents cover the cost of childcare, enabling them to work or study. The Dutch Tax Authority is responsible for assessing applications, disbursing payments, and monitoring compliance, including detecting and investigating potential fraud. This resulted in severe financial, social, and emotional distress for the affected families, including significant debt, loss of homes, unemployment, and psychological trauma (Parliamentary Interrogation Committee on Childcare Allowance, 2020).

The automated system designed to detect potential fraud in childcare allowances was introduced by the Dutch Tax Authority in 2013. From 2014 to 2017, thousands of parents began receiving unjustified accusations of fraud, resulting in demands for repayment of large sums of money. Many affected families were unable to contest these accusations effectively due to unclear and rigid bureaucratic processes. Between 2017 and 2019, media coverage and public outrage intensified as more parents came forward, prompting official investigations into the scandal. In December 2020, a parliamentary inquiry (Parliamentary Interrogation Committee on Childcare Allowance, 2020) concluded that the automated system had unjustly harmed thousands of families, highlighting severe institutional and structural failures. Subsequently, in January 2021, the Dutch government, led by Prime Minister Mark Rutte, collectively resigned, taking political responsibility for the widespread harm caused by the scandal.

Several key stakeholders played critical roles in the scandal. The Dutch Tax Authority was responsible for managing childcare benefit claims and implementing the ADM system(Parliamentary Interrogation Committee on Childcare Allowance, 2020). Parents and families were the direct victims of the wrongful fraud accusations, facing significant personal and financial hardship (Amnesty International, 2021). Public officials and the broader government faced criticism for institutional negligence, failing to adequately oversee the ADM system and insufficiently addressing the concerns of affected citizens (Autoriteit Persoonsgegevens, 2020; Parliamentary Interrogation Committee on Childcare Allowance, 2020). Regulatory bodies such as the Autoriteit Persoonsgegevens (Dutch Data Protection Authority)(2020) and human rights organizations, notably Amnesty International (2021), strongly criticized the handling of personal data and highlighted the discriminatory impacts of the ADM system.

The algorithm at the center of the scandal was originally intended to efficiently detect fraud in childcare benefit applications by automatically identifying suspicious claims (Parliamentary Interrogation Committee on Childcare Allowance, 2020). However, it heavily relied on overly simplistic and opaque indicators, lacking necessary transparency and clear mechanisms for challenging or auditing its decisions (Amnesty International, 2021). As a result, legitimate claims were systematically misidentified as fraudulent, disproportionately affecting vulnerable populations and particularly families from minority backgrounds (Amnesty International, 2021; Parliamentary Interrogation Committee on Childcare Allowance, 2020).

There are various perspectives on the potential reasons behind the failure of the Dutch childcare benefits system and who holds the primary responsibility. Some argue that the scandal was a result of structural and institutional failures, emphasizing rigid bureaucratic processes and systemic shortcomings (Parliamentary Interrogation Committee on Childcare Allowance, 2020). Others point to algorithmic bias, emphasizing that the automated system disproportionately targeted certain groups, leading to unjust outcomes (Amnesty International, 2021). Additionally, the Dutch Data

Protection Authority (Autoriteit Persoonsgegevens, 2020) criticized the tax authorities for failing to comply adequately with privacy regulations. Human rights organizations highlighted how automated decision systems could exacerbate biases, advocating for stricter guidelines and oversight (Amnesty International, 2021). Public administration experts also pointed to a significant lack of effective communication channels between affected citizens and the Tax Authority, alongside rigid internal fraud detection targets, as key contributing factors (Parliamentary Interrogation Committee on Childcare Allowance, 2020). This case is a good example of why robust accountability frameworks are essential in protecting citizens from algorithmic harms and the urgency of integrating accountability into public-sector ADM practices.

This case was chosen for analysis because it clearly shows the serious consequences that can happen when accountability is missing in algorithm-based decision-making. Although the new system was already in use, there were no proper mechanisms in place to oversee and evaluate its performance (Diakopoulos, 2016). As discussed earlier, there were different opinions about who should take responsibility and what specifically caused the scandal.

Accountability should clearly show who is responsible for certain actions and decisions. However, in the Dutch childcare benefits scandal, the lack of clear responsibility allowed systemic mistakes to go unnoticed, significantly harming thousands of people. Even though the Dutch government eventually resigned to acknowledge these serious mistakes, the damage had already occurred, highlighting the importance of having proactive accountability rather than reacting after problems arise (Bovens & Schillemans, 2014).

This case is particularly valuable for analysis because it involves policy-making, algorithmic decision-making, and administrative oversight. It shows how technology can make existing institutional weaknesses worse and emphasizes the importance of human oversight and transparency in preventing harm caused by algorithmic biases (Eubanks, 2018).

The Dutch childcare benefits scandal is not an isolated case but part of a global pattern in which algorithmic tools deployed by public institutions have produced discriminatory or harmful outcomes. Countries across the world—including China, the

United States, the United Kingdom, and Germany—are integrating ADM systems into core state functions such as policing, social benefits, credit evaluation, and immigration control (Benjamin, 2019; Yeung, 2018). Yet regulatory responses have been fragmented and inconsistent. In China, regulatory frameworks have been introduced to govern algorithmic recommendations and protect personal data, but enforcement and ethical oversight remain weak (Wang, 2024). In contrast, the European Union has advanced a rights-based regulatory approach through its AI Act, emphasizing risk classification and transparency obligations for high-stakes ADM applications (Veale & Brass, 2019; Yeung et al., 2019).

Despite these developments, many governments continue to deploy algorithmic systems without robust mechanisms for oversight, redress, or democratic accountability. As Wenzelburger et al. (2022) argue, public acceptance of ADM is shaped less by technical performance and more by the context in which it is embedded—trust in institutions, the perceived legitimacy of the decision-making process, and the availability of meaningful recourse. This highlights the need for normative frameworks that go beyond efficiency and embrace democratic values such as transparency, fairness, and the protection of fundamental rights. Studying the Dutch case through the lens of algorithmic accountability is therefore not only nationally relevant but contributes to a pressing global conversation about how to govern automated public power responsibly.

This thesis examines the Dutch childcare benefits scandal as a pivotal case study in understanding the complexities of algorithmic decision-making within public sector governance. By scrutinizing the interplay between algorithmic systems and institutional practices, this research aims to uncover how lack in fairness, transparency, and explainability contributed to systemic injustices. Through this analysis, the study seeks to derive insights that can inform the development of more accountable and equitable ADM frameworks in public administration.

To conduct this analysis, I first explored classical accountability theories. Given that Algorithmic Decision-Making is a new and rapidly evolving field, my research primarily focused on theories specifically explaining accountability within ADM, as algorithms are inherently complex and require distinct accountability mechanisms. I mainly examined three key principles—Fairness, Transparency, and Explainability—as proposed by Diakopoulos’(2016) Algorithmic Accountability Framework. The literature review section will further elaborate on both the case and these accountability theories.

The research methodology was qualitative, as this approach effectively captures the complexities involved in analyzing accountability. Data collection involved conducting interviews with various stakeholders and examining secondary sources, especially online interviews with public officials involved in the case.

The thesis aims to conduct a case study analysis of accountability mechanisms in the Dutch childcare benefits scandal. It will evaluate how effectively the Dutch childcare benefits algorithm adhered to Diakopoulos' principles of Fairness, Transparency, and Explainability. Furthermore, the thesis will propose actionable strategies to better integrate these algorithmic accountability principles into public-sector decision-making systems. The central research questions are: Why and to what extent did the Dutch childcare benefits algorithm fail to meet Diakopoulos’ principles? And what lessons can be learned from this case to enhance algorithmic accountability in public decision-making.

2. Literature Review

Since the early 2000s, and particularly throughout the 2010s, automated decision-making systems have become increasingly common in public administration, being widely adopted in areas like welfare management, taxation, immigration, healthcare, and law enforcement. This rapid growth has sparked significant interest from scholars, policymakers, and the general public due to the serious implications ADM can have for fairness, transparency, and accountability (Eubanks, 2018; Yeung, Howes & Pogrebna, 2019).

Although there is already a substantial body of academic literature and policy discussion about ADM, the field continues to evolve quickly, driven by ongoing technological advances. Policymakers often find it challenging to regulate ADM effectively because new technologies emerge rapidly, methods for implementation differ greatly, and regulatory environments constantly shift (Diakopoulos, 2019; Barocas, Hardt & Narayanan, 2019). Researchers emphasize that these challenges make it difficult for public officials to ensure ADM systems are fair, transparent, and free from bias or discrimination (Pasquale, 2015; O'Neil, 2016).

This literature review focuses specifically on how the principles of algorithmic accountability—fairness, transparency, and explainability—have been understood and debated in scholarly work. In particular, it critically engages with Nicholas Diakopoulos's Algorithmic Accountability Framework (Diakopoulos, 2015; 2019; 2020) and compares it to other important accountability theories, such as those developed by Bovens (2007) and Bovens and Schillemans (2014). By exploring these perspectives, this review aims to provide context for analyzing the shortcomings of ADM accountability mechanisms, illustrated through the example of the Dutch childcare benefits scandal.

Understanding algorithmic accountability requires revisiting broader definitions of accountability in public administration and examining how these ideas are adapted in the context of automated decision-making. Traditional views on accountability, such as the framework developed by Mark Bovens (2007), define it as a relationship in which an actor must explain and justify their conduct to a forum, which then assesses the explanation and may impose consequences. This model has been widely used in public administration but faces limitations when applied to ADM systems, which are often complex, opaque, and involve multiple actors across institutional boundaries.

Bovens and Schillemans (2014) later refined this approach by distinguishing between legal, political, and professional accountability. They highlighted how accountability mechanisms can become fragmented in complex governance environments, making it difficult to assign clear responsibility. These insights are particularly relevant for ADM, where algorithms are often embedded in bureaucratic systems, and outcomes are shaped not just by code, but by organizational routines, data infrastructures, and political goals.

To address these gaps, Nicholas Diakopoulos has proposed a more targeted framework for algorithmic accountability. Originally developed in the context of investigative journalism (Diakopoulos, 2015), his work has expanded to encompass broader governance and policy contexts (Diakopoulos, 2019). His Algorithmic Accountability Framework emphasizes several key principles: fairness, transparency, explainability, auditability, redressability, responsibility, and accuracy. These principles aim to provide a structured approach to evaluating how ADM systems function and whether they can be held accountable when things go wrong.

Diakopoulos's approach is particularly useful because it directly engages with the challenges posed by algorithmic opacity and complexity. He argues that traditional accountability models need to be adapted to deal with black-box systems, limited access to proprietary algorithms, and the technical barriers that prevent meaningful oversight. In contrast to more abstract frameworks, Diakopoulos offers practical guidance—such as using algorithmic impact assessments and audits—to operationalize accountability in real-world contexts (Diakopoulos & Friedler, 2021).

This conceptual foundation is important because it helps clarify how different approaches to accountability apply in the context of ADM. Traditional models, like those proposed by Bovens (2007, 2014), emphasize formal reporting relationships within institutions, which work well in relatively transparent, hierarchical settings. However, ADM introduces new challenges—such as distributed responsibility, technical opacity, and proprietary algorithms—that these traditional models were not designed to handle. In contrast, algorithmic-specific frameworks, such as Diakopoulos's, directly address these issues by proposing new principles and tools tailored for evaluating how decisions are made by or with the help of algorithms. Understanding the strengths and limitations of both approaches is essential for building accountability systems that can function effectively in public-sector contexts where ADM is increasingly used.

2.2 Transparency and the Limits of Openness

Transparency is frequently presented as one of the most essential conditions for algorithmic accountability. However, while scholars widely agree on its importance, they differ in how it should be implemented and what it can realistically achieve. Diakopoulos (2015, 2020) treats transparency as a means of enabling public oversight and institutional accountability, particularly in contexts where decision-making systems are opaque. His proposals—such as algorithmic impact assessments and structured disclosures—are intended to make algorithmic processes more understandable to journalists, policymakers, and the public.

Pasquale (2015, pp. 3-5) offers a more critical view. While Diakopoulos sees transparency as a practical tool that can support investigative scrutiny and policy reform when structured disclosures are provided, Pasquale is more skeptical about the power of transparency in practice. He argues that ADM systems often function as "black boxes," meaning that the internal logic, inputs, and decision rules of these systems are hidden from public view. This opacity, he suggests, allows powerful actors—such as corporations, government agencies, and data brokers—to avoid scrutiny and accountability for the decisions these systems make. In fields like finance and healthcare, algorithms may determine creditworthiness or insurance eligibility based on proprietary models that individuals cannot access or challenge. In contrast to Diakopoulos's more optimistic, solution-oriented view, Pasquale (2015, pp 115-116) emphasizes that the very structure of ADM systems often reinforces power imbalances and perpetuates inequalities. He warns that unless there are strong legal frameworks and institutional mechanisms to open up these systems, transparency may remain superficial and ineffective. Similarly, Ananny and Crawford (2018, 978-980) caution against assuming that transparency alone can guarantee accountability. In their view, simply revealing information about how an algorithm works does not necessarily empower affected individuals or improve institutional practices. They argue that transparency, especially when poorly defined or implemented, can create a false sense of openness—what they refer to as the illusion of accountability. For example, releasing technical documentation or model summaries might satisfy formal disclosure requirements without making the system understandable or actionable to users, auditors, or regulators. In some cases, this kind of symbolic transparency can even obscure deeper structural issues by suggesting that oversight exists when it does not. Their critique aligns with Pasquale's skepticism and pushes the conversation further by highlighting that meaningful transparency requires active interpretation, public engagement, and institutional frameworks that support scrutiny and response.

These concerns are echoed in more recent work by Diakopoulos and Friedler (2021), who emphasize that transparency must be context-specific. They outline several dimensions of transparency—such as data transparency, process transparency, and outcome transparency—and argue that each must be addressed for ADM systems to be truly accountable. For example, stakeholders need to know what data are used and how they are collected and processed (data transparency); they also need information about the design logic and operational parameters of the algorithm (process transparency); and finally, they must understand the consequences and patterns in the decisions being made (outcome transparency).

Despite this layered view of transparency, many challenges remain. Legal and institutional barriers, including intellectual property protections and privacy concerns, often limit access to algorithmic systems. Moreover, even when transparency is technically achieved, the information disclosed may be too complex or abstract to be meaningful to most stakeholders (Ananny & Crawford, 2018, pp. 979–980; Diakopoulos, 2016, p. 58). As a result, scholars argue that transparency should not be seen as a standalone solution but must be embedded in broader institutional practices—such as independent audits, regulatory oversight, and mechanisms for public contestation (Ananny & Crawford, 2018; Kroll et al., 2017).

Other real-world examples also demonstrate the limitations of transparency in ADM systems. For instance, predictive algorithms used in the U.S. criminal justice system, such as Correctional Offender Management Profiling for Alternative Sanctions

(COMPAS), have faced significant criticism for being opaque and difficult to audit (Angwin et al., 2016; Rudin, 2019). Despite being used to inform bail and sentencing decisions, the workings of COMPAS was not transparent, and affected individuals are often unaware of how their risk scores are calculated. This lack of transparency has made it nearly impossible to contest potentially unfair outcomes and has raised serious concerns among civil liberties organizations (Larson et al., 2016). Such cases show how technical opacity can lead to real-world consequences and reinforce structural inequalities.

2.3 Fairness in Algorithmic Decision-Making

Fairness is one of the widely discussed yet conceptually complex principles in algorithmic accountability. In the context of ADM, fairness generally refers to the extent to which algorithms treat individuals and groups equitably, without causing unjustified bias or discrimination. However, scholars differ significantly in how fairness is defined, measured, and operationalized.

Barocas, Hardt, and Narayanan (2019) offer one of the most influential frameworks for understanding fairness in machine learning systems. They emphasize that fairness is not a single standard but rather a family of sometimes conflicting criteria, including demographic parity, equalized odds, and individual fairness. Demographic parity focuses on ensuring equal outcomes across demographic groups, while equalized odds accounts for both false positives and false negatives across groups. Individual fairness, in contrast, holds that similar individuals should receive similar outcomes. Barocas et al. also highlight that it is rarely possible to satisfy all definitions of fairness simultaneously, and that fairness assessments must be context-specific. Their work is largely theoretical but supported by examples from algorithmic risk assessment, hiring tools, and credit scoring, where different fairness metrics can lead to contradictory conclusions. For instance, optimizing for equalized odds may worsen disparities in demographic parity, forcing policymakers and system designers to make trade-offs depending on the specific goals and social values relevant to the application context.

The complexity of fairness is particularly evident in ADM systems trained on historical data, which can reflect and amplify past societal biases. O’Neil (2016), provides the example of predictive policing algorithms in the United States that disproportionately target minority neighborhoods. These systems rely on historical arrest data, which itself reflects racially biased policing patterns. As a result, the algorithms produce outputs that continue to over-police the same communities, reinforcing systemic discrimination under the guise of neutral data analysis. Similarly, Eubanks (2018), explores how welfare eligibility systems in Indiana and Pennsylvania used rigid automation to manage public assistance programs. These systems often failed to account for the lived realities of poor families, applying rules mechanically and terminating benefits without meaningful human oversight. Her ethnographic research highlights how these systems dehumanized welfare recipients and undermined their access to critical resources, all while presenting themselves as neutral and efficient technologies.

Diakopoulos (2019) integrates fairness into his algorithmic accountability framework as a core component, arguing that fairness is essential for legitimacy and public trust. He stresses the importance of designing systems that are not only technically accurate but also socially responsible. This includes auditing datasets for bias, consulting with affected communities, and ensuring that fairness considerations are prioritized throughout the design and implementation process. Diakopoulos acknowledges the technical challenges of measuring fairness, he sees it as a practical design goal that can be promoted through deliberate governance strategies.

Other scholars, like Zarsky (2016), approach fairness from a regulatory and legal standpoint, emphasizing that ADM cannot be assessed solely on technical grounds. Zarsky argues that the promise of algorithmic efficiency often obscures its potential to institutionalize unfairness, especially when algorithms rely on proxy variables—such as ZIP codes or internet usage patterns—that correlate with race, income, or other protected characteristics. He suggests that even when the data used are legally obtained and statistically valid, the outcomes can still be discriminatory. Zarsky calls for a robust legal and regulatory infrastructure that includes regular audits, transparency mandates, and avenues for redress. He also emphasizes the importance of public accountability mechanisms to ensure that algorithmic tools do not circumvent anti-discrimination laws or ethical standards.

What unites these perspectives is the recognition that fairness cannot be reduced to technical metrics alone. Fairness in ADM is inherently socio-political and value-laden. As Benjamin (2019) argues, many algorithmic systems embed and reproduce racial and social hierarchies under a veneer of objectivity. To ensure fairness, he argues, ADM must be critically examined through the lens of structural inequality, and efforts to make systems fair must go beyond data cleaning or technical tweaks to include meaningful social and institutional reform. Together, these contributions highlight that fairness in ADM is a multidimensional challenge that requires both technical solutions and broader social and policy interventions.

2.4 Explainability and the Black Box Problem

As algorithmic decision-making systems grow more complex and influential in public administration (Diakopoulos, 2016), the need for explainability becomes increasingly important. While transparency refers to making systems visible or open in some way—such as by disclosing data sources, code, or policy rationales—explainability focuses on whether the outputs of a system can be understood and interpreted by humans. In other words, explainability is not just about access to information but about making that information meaningful and actionable for those affected by algorithmic decisions (Diakopoulos, 2015).

Diakopoulos (2015, pp 5-6) identifies explainability as a key principle for algorithmic accountability. He argues that for accountability mechanisms to function properly, individuals must be able to understand how and why a decision was made. This is particularly important in the public sector, where government decisions can deeply affect people's rights and access to services. Diakopoulos (2015, p 12) highlights that explainability is not only a technical issue but also a communicative one—it depends on how well an explanation is tailored to the audience, whether it be developers, civil servants, or members of the public.

This raises the question: explainability for whom? Technical documentation may be sufficient for engineers or data scientists but useless for policymakers, legal professionals, or citizens without specialized training. A decision that is explainable to a developer may remain a mystery to a layperson, especially if the explanation is filled with mathematical jargon or relies on complex statistical reasoning. As a result, many scholars argue that explainability must be context-sensitive and audience-specific. Meaningful explainability should empower not only internal stakeholders who manage or design ADM systems but also those who are affected by these systems—including the public (Weller, 2019; Lipton, 2018).

Eubanks (2018) takes this argument further by demonstrating how the lack of explainability in public sector ADM can have real-life, harmful consequences. She describes cases in which families were denied access to welfare programs by systems that could not provide clear reasons for their decisions. Because the people affected by these decisions did not understand how or why they were flagged, they had no way to contest the decisions or appeal them effectively. Eubanks argues that this opacity undermines both fairness and democratic accountability. It disproportionately harms marginalized communities, who often lack the legal or technical resources to navigate these opaque systems.

The lack of explainability also limits the legitimacy of ADM systems. Public legitimacy depends on the ability of individuals to understand and challenge decisions that affect their lives (Diakopoulos, 2016). Without clear explanations, trust in institutions may erode, particularly when algorithmic outcomes are perceived as arbitrary or discriminatory. Explainability, therefore, plays a crucial role not only in accountability but also in supporting contestability—the ability to dispute, correct, or reverse decisions (Doshi-Velez & Kim, 2017). A system that cannot explain its outcomes is one that cannot be effectively governed or held responsible. Explainability is closely linked to but distinct from transparency. While transparency opens the system, explainability interprets it. For ADM systems to be legitimate and accountable, they must not only be transparent to auditors and developers but also explainable to the people affected by their outcomes. Without this, public trust and democratic oversight are compromised.

2.5 The Role of Institutions and Bureaucracy in Algorithmic Accountability

While much of the literature on algorithmic accountability focuses on technical solutions—such as explainability, transparency, or fairness metrics—scholars in public administration emphasize that institutional and bureaucratic contexts are equally important. ADM systems do not operate in isolation. They are embedded within complex bureaucracies that shape how algorithms are designed, implemented, and overseen. These institutional settings influence how accountability is understood, practiced, and enforced.

Romzek and Dubnick (1987) emphasize that accountability is shaped by the organizational culture and bureaucratic norms within public institutions. They argue that different types of accountability—hierarchical, legal, political, and professional—may compete or conflict within agencies. For instance, performance-based accountability mechanisms might reward efficiency and cost-cutting, even if they compromise fairness or public service values. In algorithmic systems, this could lead agencies to adopt risk-scoring tools or automation systems that prioritize detecting fraud over protecting rights—especially when institutional incentives are focused on minimizing financial losses rather than serving vulnerable populations.

Similarly, Hood (2010, pp 990 - 991) and Schillemans (2011, pp 388-389) point out that accountability is not just a technical design problem but also a political and institutional challenge. Bureaucracies often face competing pressures: political actors may push for stricter enforcement, civil servants may operate under performance targets, and regulatory frameworks may lag behind technological developments. These dynamics can lead to what some scholars describe as “accountability theater”—formal mechanisms that appear to provide oversight without addressing deeper power imbalances or systemic flaws (Schillemans, 2011).

The literature also suggests that institutional self-protection and blame avoidance strategies may weaken real accountability. When ADM systems produce harmful outcomes, public agencies may attribute the decisions to the algorithm itself—describing it as an objective or technical process—and thereby deflect responsibility. This depersonalization of decision-making can obscure who is actually accountable and create barriers for those trying to contest or appeal decisions (Dubnick, 2005; Ananny & Crawford, 2018).

These insights highlight why algorithmic accountability cannot be fully addressed through technical fixes alone. Institutions must create governance structures that establish clear lines of responsibility, support ethical decision-making, and provide effective avenues for redress. This requires not only technical audits but also institutional audits—scrutinizing how public agencies integrate ADM into their operations, what values are prioritized, and how accountability mechanisms are institutionalized. Without considering the wider organizational environment—its goals, incentives, constraints, and power relations—attempts to make ADM systems more accountable will likely fall short. Scholars from computer science, law, public administration, and critical data studies converge on the view that ADM systems raise serious concerns about fairness, transparency, and explainability—particularly in public-sector applications where decisions can have life-altering consequences.

A recurring theme is the inadequacy of traditional institutional accountability models when applied to ADM. Frameworks like those of Bovens (2007) and Bovens & Schillemans (2014) remain valuable for understanding bureaucratic dynamics and institutional responsibility but fall short in capturing the technical opacity. Meanwhile, technical approaches tend to prioritize transparency or explainability, but often fail to account for the real-world organizational and political contexts in which these systems operate (Pasquale, 2015; Ananny & Crawford, 2018).

This literature also reveals unresolved tensions. For example, while fairness is a widely endorsed goal, there is no single agreed-upon metric for assessing it. Definitions such as demographic parity, equalized odds, and individual fairness often conflict in practice, requiring policymakers to make trade-offs that are as much ethical and political as they are technical (Barocas, Hardt & Narayanan, 2019). Similarly, explainability is widely regarded as essential for contestability and public trust, yet scholars disagree on what constitutes a meaningful explanation—and for whom (Lipton, 2018; Diakopoulos, 2015). These tensions reflect deeper structural challenges about how to govern ADM in a way that balances efficiency, fairness, and democratic legitimacy.

Given these complexities, this thesis adopts Diakopoulos' Algorithmic Accountability Framework as its analytical lens. Diakopoulos brings together technical, ethical, and institutional dimensions of accountability and explicitly addresses the challenges of fairness, transparency, and explainability in automated systems. His framework offers both normative guidance and practical tools—such as impact assessments, structured audits, and stakeholder-oriented explanations—that can be used to evaluate real-world ADM systems. Unlike more abstract or purely theoretical models, Diakopoulos's approach is designed to be actionable and adaptable to public-sector contexts. This makes it particularly suitable for analyzing how algorithmic accountability can (or fails to) function in complex governance environments.

The Dutch childcare benefits scandal represents a powerful and highly relevant case through which to explore and test these concepts. It is one of the most well-documented examples of how an ADM system—designed to detect fraud in welfare programs—produced unjust and discriminatory outcomes. What makes the Dutch case especially compelling is that the algorithm operated within a democratic state with strong legal institutions and public accountability mechanisms. Despite this, the ADM system remained largely unaccountable to the public, and its decisions were neither explainable nor contestable to the affected families or even oversight bodies (Amnesty International, 2021). This paradox—where a system designed for fairness and efficiency ended up undermining both—makes the Dutch case a meaningful setting for evaluating the effectiveness and limitations of algorithmic accountability frameworks. By applying Diakopoulos's framework to the Dutch case, this study aims to assess how well the system adhered to the principles of fairness, transparency, and explainability.

3. Theoretical Framework

This section outlines the analytical framework used to analyze the Dutch childcare benefits scandal. It combines two complementary approaches: Nicholas Diakopoulos's Algorithmic Accountability Framework and the broader concept of algorithmic governance. While Diakopoulos's framework provides a set of normative criteria to assess whether automated decision-making systems uphold accountability (Diakopoulos, 2015; 2019), algorithmic governance offers a broader explanation of how and why such systems are implemented within public institutions (Yeung, 2018; Ziewitz, 2016). By integrating these approaches, this thesis aims to both evaluate and contextualize the accountability failures associated with the Dutch case.

The use of algorithms in public administration presents both opportunities and challenges for governance, particularly in terms of accountability, transparency, and fairness. As ADM systems increasingly influence decisions in areas such as taxation, welfare, and immigration, questions have emerged about how these systems can be held accountable when they produce harmful or discriminatory outcomes (Eubanks, 2018; Yeung, 2018). This thesis adopts a dual analytical framework to address these questions. It uses Diakopoulos's principles of algorithmic accountability as the main analytical tool to assess how the Dutch childcare benefits algorithm performed with regard to fairness, transparency, and explainability. In addition, it draws on the concept of algorithmic governance to understand the broader structural and institutional dynamics that led to the system's deployment and contributed to its lack of accountability (Ziewitz, 2016).

Diakopoulos's framework, which has already been discussed in detail in the literature review, provides clear normative standards against which ADM systems can be evaluated. These standards inform the core research questions of this thesis and serve as the basis for designing the interview guide and structuring the case study analysis. At the same time, the concept of algorithmic governance complements this approach by highlighting the systemic factors—such as the pursuit of efficiency, automation of bureaucratic discretion, and belief in data-driven objectivity—that influence how ADM systems are embedded in public institutions (Yeung, 2018).

Together, these two frameworks provide a comprehensive lens for examining both the design and the broader political-institutional context of the Dutch case.

3.2 Algorithmic Accountability Framework

The Algorithmic Accountability Framework developed by Nicholas Diakopoulos serves as the primary analytical tool for this thesis. Although already outlined in the literature review, its role here is to guide the case analysis by offering concrete criteria against which the performance of the Dutch ADM system can be assessed. Diakopoulos's framework is rooted in the idea that algorithms, especially those used in public decision-making, must be subjected to scrutiny similar to traditional forms of governance (Diakopoulos, 2015; Diakopoulos & Friedler, 2021). His framework emphasizes a set of principles—including fairness, transparency, and explainability—that serve as benchmarks for evaluating whether an ADM system upholds democratic accountability.

Fairness refers to the degree to which algorithmic systems treat individuals and groups impartially, without unjustified discrimination. This includes both procedural fairness—ensuring that the decision-making process itself is free from bias—and distributive fairness—ensuring that outcomes are equitably distributed across different demographic groups. A fair ADM system should not replicate or amplify existing social biases embedded in historical data or decision logic (Diakopoulos & Friedler, 2021).

Transparency is concerned with the visibility of algorithmic processes. It involves disclosing how data are collected, what models are used, and how decisions are made. However, transparency is not just about making information available—it must be provided in a meaningful, understandable form that enables stakeholders, including citizens and oversight bodies, to comprehend and evaluate how decisions are reached (Diakopoulos, 2016, pp. 57–59).

Explainability, while closely related to transparency, focuses specifically on the ability to articulate and justify individual algorithmic decisions in a way that is accessible to non-technical stakeholders. Explainability is crucial for enabling those affected by decisions to understand why a particular outcome occurred, and it supports contestability by giving individuals the means to challenge unjust decisions (Diakopoulos, 2015, pp. 5–6, 12).

In this study, these principles are used not only to assess the system's design but also to interrogate its outcomes and the institutional responses to those outcomes. Each principle is operationalized into specific questions within the interview guide, allowing for a structured assessment of whether the algorithm treated individuals equitably (fairness), allowed for scrutiny and understanding (transparency), and offered sufficient explanations to affected individuals and oversight bodies (explainability). By applying this framework to the Dutch childcare benefits scandal, the thesis seeks to offer a grounded, principle-based evaluation of algorithmic accountability in practice.

3.3 Algorithmic Governance

The concept of algorithmic governance provides the second pillar of this thesis's analytical framework. It offers a broader perspective for understanding how and why ADM systems are introduced, legitimated, and embedded into public-sector institutions. While Diakopoulos's framework is primarily normative and evaluative, algorithmic governance emphasizes the socio-technical, political, and institutional dynamics that shape ADM deployment.

Algorithmic governance refers to the use of algorithms and data-driven systems as instruments for governing populations and managing public services. Scholars such as Yeung (2018) and Ziewitz (2016) argue that ADM technologies do not exist in a vacuum; they are embedded in bureaucratic cultures, political ideologies, and institutional arrangements that shape how they function and what kinds of risks they produce. These systems often reflect underlying logics of efficiency, risk management, and control, which can shift accountability away from human decision-makers and obscure the normative implications of automated decisions. These risks are particularly evident under specific institutional conditions, as explored in more empirical works. For example, Eubanks (2018) documents how ADM systems in U.S. welfare and child protection programs amplify structural inequalities when implemented in contexts driven by performance metrics. Similarly, Amnesty International (2021) illustrates how prioritizing fraud detection in public benefits systems without proper oversight can lead to discriminatory outcomes.

One key concern in algorithmic governance is the diffusion of responsibility. ADM systems often involve multiple actors—data scientists, software developers, policy designers, and frontline administrators—each playing a partial role in shaping how the system functions. As a result, it can become difficult to identify who is accountable when things go wrong (Ananny & Crawford, 2018). Moreover, technical complexity and legal constraints—such as proprietary software or data protection laws—may limit public insight into how these systems operate (Pasquale, 2015).

Another central insight of the algorithmic governance literature is that automated decision-making systems may not simply optimize or neutralize administrative processes, but rather reinforce or entrench existing power structures. Algorithms, far from being objective tools, can institutionalize specific policy preferences, reproduce historical biases, and disproportionately target already marginalized populations (Eubanks, 2018). This concern is particularly relevant in the public sector, where citizens affected by automated decisions often lack the resources, knowledge, or legal support to effectively challenge them. Eubanks (2018), through detailed case studies in the United States, explores the specific conditions under which ADM systems produce and intensify inequality. Her analysis shows that when such systems are implemented in contexts already marked by structural disadvantage—such as poverty, racial discrimination, or bureaucratic opacity—they often amplify existing injustices rather than mitigate them. For example, she highlights how data-driven systems used in welfare eligibility or child protection prioritize efficiency and risk management over fairness and social support. These systems tend to penalize individuals who already face social and economic vulnerabilities, often without providing clear pathways for appeal or human intervention.

The integration of Diakopoulos's Algorithmic Accountability Framework and the concept of algorithmic governance enables a multi-dimensional analysis of the Dutch childcare benefits scandal. While Diakopoulos's framework focuses on if the accountability principles were met —specifically, whether the ADM system upheld principles of fairness, transparency, and explainability—algorithmic governance provides insights into why it the principles were not held by examining the institutional logic, bureaucratic incentives, and systemic conditions under which the ADM system was developed and implemented.

This dual approach allows for a more complete understanding of the case. Diakopoulos's principles guide the evaluation of the algorithm's design and its real-world outcomes, offering a structured basis for identifying failures in accountability. At the same time, the concept of algorithmic governance contextualizes these failures within broader administrative and political practices that may have prioritized efficiency, risk aversion, or fraud detection over citizen rights and democratic accountability.

4. Methodology

This study employs a qualitative research approach to explore to what extent the Dutch childcare benefits algorithm adhered to the algorithmic accountability principles of Explainability, Transparency, and Fairness as proposed by Diakopoulos (2019) and from the perspective of algorithmic governance, explore why they were not adhered in practice. A qualitative approach is suited for examining complex social phenomena like algorithmic accountability because it allows for in-depth exploration of perceptions, institutional contexts, and interpretive meanings behind algorithmic practices (Denzin & Lincoln, 2011; Yanow & Schwartz-Shea, 2014).

Specifically, this research seeks to examine how the Dutch childcare benefits algorithm is aligned with the principles of algorithmic accountability by focusing on three main aspects. First, it investigates whether the algorithmic decision-making process was sufficiently explainable to all relevant stakeholders, including affected families, legal professionals, journalists, and policymakers.

Second, the study assesses the extent to which the criteria and processes behind the ADM system were transparent. This includes evaluating whether clear and accessible information was available not only to the public but also to institutional actors responsible for oversight and accountability.

Third, the research explores whether the algorithm systematically disadvantaged certain demographic or socio-economic groups and whether appropriate safeguards were in place to detect, prevent, or mitigate unfair outcomes. This involves analyzing both the design and implementation of the ADM system, as well as the institutional responses to early signs of bias or disproportionate harm. In addition, the research aims to understand why these accountability principles were not upheld in the Dutch case.

To achieve these aims, the study combines primary data collection through semi-structured qualitative interviews with secondary data analysis. Publicly available reports, parliamentary documents, and expert commentaries were used to triangulate findings, validate interpretations, and provide a richer understanding of how algorithmic accountability was—or was not—achieved in this case.

4.2 Data Sources

In addition to conducting primary interviews, the study draws on a range of secondary data sources to strengthen the validity and comprehensiveness of the analysis. These include:

- Recorded interviews and testimonies from victims, politicians, lawyers, and civil servants—publicly available via media outlets, podcasts, and investigative journalism platforms.
- Documentaries and films on the scandal, such as the Dutch movie "Alleen tegen de Staat" (2023), provide in-depth personal narratives of affected families and expert commentary.
- Parliamentary hearings, Parliamentary Interrogation Committee on Childcare Allowance reports (2020), and official investigations into the misuse of ADM by the Dutch tax authorities.
- Academic and NGO reports, such as Amnesty International's Xenophobic Machines (2021), which offer critical perspectives on algorithmic discrimination.

These secondary sources supplemented the primary interviews and also helped validate patterns and themes across different types of evidence.

4.3 Participant Recruitment

Participants for the primary interview component were selected purposefully to ensure diverse representation from key stakeholder groups, including: a victim directly affected by the ADM system; journalists who investigated or reported on the scandal; a legal expert who represented victims and analyzed the case; and specialists in artificial intelligence and machine learning with expertise in ADM.

Recruitment occurred through direct outreach, snowball sampling, and engagement in professional or thematic online communities in Reddit and Facebook. 8 participants were interviewed to capture a range of views and experiences.

4.4 Data Collection Methods

Primary data were collected through semi-structured online interviews (see Annex 1 for interview guide). Thematic flexibility allows exploration of individual experiences while ensuring consistency across key areas aligned with Diakopoulos's framework: explainability, transparency, and fairness.

Secondary data (documentaries, reports, public interviews) were systematically reviewed and analyzed using the same thematic framework to enable direct comparison and integration with primary interview findings.

All interviews were transcribed and coded for analysis. For publicly available sources, only ethically accessible materials were used, intended for public use, and do not infringe on individuals' privacy or confidentiality and proper citations and acknowledgments were made.

4.5 Data Analysis

Data from both primary and secondary sources were analyzed using thematic analysis, based on the six-step framework by Braun and Clarke (2013):

- Familiarization with the data
- Generating initial codes
- Searching for themes
- Reviewing themes
- Defining and naming themes
- Producing the final report

Themes were structured around the accountability principles of fairness, transparency, and explainability. Coding was facilitated through tools like Microsoft Excel and Google Sheets, ensuring consistency in handling both interview transcripts and secondary materials.

4.6 Ethical Considerations

For the primary data collection, informed consent, confidentiality, and anonymity were ensured. Participants were anonymized using pseudonyms and role descriptors.

For secondary sources, only publicly available content was used. Where individual testimonies were cited from media or documentaries, proper acknowledgment and ethical standards were maintained to respect the rights and dignity of those individuals.

4.7 Methodological Limitations

Access limitations restricted the recruitment of certain stakeholders, such as government officials or system designers, which could result in gaps in institutional perspectives. Additionally, self-selection bias may influence the range of viewpoints captured in interviews, as individuals who choose to participate may have particularly strong experiences or opinions. The secondary data used in this study also come with limitations, including potential biases in how public narratives are framed or edited, which may shape interpretation. Nevertheless, the research maintains methodological rigor through triangulation of sources, careful design of interview questions, and adherence to ethical standards. As such, it remains well-positioned to generate meaningful insights into the Dutch childcare benefits algorithm and its adherence to critical accountability principles.

5. Case study analysis

This chapter presents the core empirical analysis of the Dutch childcare benefits scandal by systematically applying Diakopoulos' Algorithmic Accountability Framework and the concept of algorithmic governance. The aim is to assess how and to what extent the childcare benefits algorithm violated the principles of fairness, transparency, and explainability, as outlined in the research question. The analysis draws on a qualitative case study methodology combining primary and secondary data sources. In addition to document analysis of official reports, legal inquiries, the study integrates original data from eight semi-structured interviews. These included one victim directly affected by the childcare benefits sanctions, two investigative journalists involved in exposing the case, one legal expert and four experts in ADM from the Netherlands and Armenia. The interviews were conducted between March and May 2025 and thematically coded to identify recurring patterns and perceptions across the three pillars of accountability. By triangulating these diverse data sources, the analysis provides a comprehensive, multi-actor perspective on the systemic accountability failures that characterized this landmark case of algorithmic decision-making in public administration.

The analysis proceeds by examining each of the three pillars of Diakopoulos' framework in turn, beginning with fairness as the foundational principle whose violation most directly contributed to the disproportionate harms experienced by affected citizens.

5.1 Fairness

The principle of fairness in algorithmic accountability refers to the commitment that automated systems must not replicate, exacerbate, or institutionalize unjust treatment, particularly toward vulnerable populations (Barocas, Hardt, & Narayanan, 2019). Diakopoulos frames fairness as both a procedural and distributive obligation—requiring that the processes guiding algorithmic decision-making are just, and that the outcomes they produce do not disproportionately disadvantage specific social groups. The Dutch childcare benefits scandal represents a paradigmatic failure of fairness on both dimensions: the algorithm's outputs were discriminatory, its design and deployment neglected questions of justice, and its institutional environment prioritized efficiency over equity (Autoriteit Persoonsgegevens, 2020).

The childcare benefits algorithm—officially termed a "risk classification model" (Autoriteit Persoonsgegevens, 2020)—used self-learning techniques to identify families likely to commit fraud in their childcare benefit applications. Although nationality or ethnicity was not directly encoded as a variable, features such as dual nationality, foreign-sounding names, residence in certain postal codes, and language used in correspondence served as proxies for racial and ethnic identity (Amnesty International, 2021; Autoriteit Persoonsgegevens, 2020). One algorithmic expert explained:

> "When you select input variables that strongly correlate with ethnicity, and then use them in an opaque way to trigger investigations, you are building systemic bias into the machine."

These design choices were not value neutral. As the Amnesty International report (2021) notes, parents with migration backgrounds were significantly more likely to be flagged for investigation. Specifically, individuals of Turkish and Moroccan descent were among the most disproportionately affected—amounting to a system of institutionalized ethnic profiling (Autoriteit Persoonsgegevens, 2020). The algorithm's reliance on these correlated variables resulted in indirect discrimination (Amnesty International, 2021; Autoriteit Persoonsgegevens, 2020), a concept well-established in both human rights law and machine learning ethics: a formally neutral criterion produces systematically harmful effects on protected groups (Barocas, Hardt, & Narayanan, 2019). A journalist who had investigated the scandal echoed this view, stating:

"We kept seeing the same pattern: it was people with dual nationality, people whose Dutch wasn't fluent, people who looked 'foreign.' The state built an ethnic filter into its welfare system and pretended it was neutral."

From the standpoint of computational ethics, the model violated both individual fairness (treating similar individuals differently based on irrelevant characteristics) and group fairness (systematic disparities in treatment between groups). The tax authority's algorithm flagged families not based on evidence of actual fraud but on features statistically associated with higher "risk"—without any individualized justification (Autoriteit Persoonsgegevens, 2020).

One victim interviewed in the study by Konaté and Pali (2023) recounted the long-term physical and psychological consequences of the wrongful fraud accusation:

"When I was fired [as a result of the wrongful accusation of fraud], I lost 34 kg in three months. As a result, I got a heart attack and part of my body became sort of paralysed. What I have is a conversion disorder. If I experience stress, this side of my body just falls out [...] It is just so heavy [referring to the 'childcare-benefit scandal']. I dyed my hair two and a half weeks ago and look, I'm grey again already. None of this is good or healthy, I cannot recover." (Konaté & Pali, 2023, p. 146)

These experiences reflect more than individual hardship; they indicate a broader pattern of racialized suspicion embedded into policy through automation. The sense of being targeted not because of actions but because of identity was a common theme across victims. As the victim I interviewed stated:

"It felt like I was being punished for who I am, not what I did."

This violated a central tenet of modern welfare provision: that services be allocated based on verified need and entitlement, not statistical suspicion. The rigid and uniform application of rules—such as requiring original receipts for every childcare expense—failed to account for individual circumstances (Parliamentary Interrogation Committee on Childcare Allowance, 2020). Victims who faced language barriers often struggled to understand official correspondence or complete complex administrative procedures. Those with limited digital literacy had difficulty accessing or navigating online systems required for compliance (Amnesty International, 2021). Families who arranged childcare informally, especially within migrant or low-income communities, lacked formal documentation and were therefore unable to prove their entitlement (Amnesty International, 2021).

The discriminatory impact of the algorithm cannot be reduced to ethnicity alone. Many affected parents were low-income, single mothers, or recent migrants with limited institutional knowledge and legal resources. These intersecting identities magnified their vulnerability. A parent with dual nationality and no Dutch-speaking support network, for example, would have faced not only an increased risk score but significant difficulty contesting the decision once flagged (Autoriteit Persoonsgegevens, 2020; Amnesty International, 2021).

The welfare algorithm thus failed to reflect a relational or equity-oriented notion of fairness. It ignored differences in social position, historical disadvantage, and the growing effect of state scrutiny—an especially strong omission for a policy domain meant to support families and reduce child poverty. From 2012 onward, Dutch welfare policy shifted toward a zero-tolerance approach to fraud, with intensified surveillance measures and aggressive enforcement (Parliamentary Interrogation Committee on Childcare Allowance, 2020, pp. 12–15). The childcare benefits system increasingly relied on algorithmic profiling to detect “risk” rather than actual wrongdoing (Autoriteit Persoonsgegevens, 2020; Amnesty International, 2021).

Fairness was not an explicit consideration at any stage of the risk classification model’s life cycle. As documented by the Parliamentary Interrogation Committee on Childcare Allowance (2020) and further analyzed by Amnesty International (2021), the algorithm was developed and deployed without any algorithmic impact assessment or structured evaluation of fairness risks. There was no internal review of potential disparate impacts, despite clear indications that certain groups such as families with dual nationality were disproportionately affected. Furthermore, no external audit mechanisms were in place to evaluate the model's legality, proportionality, or discriminatory outcomes. The process also lacked any form of participatory oversight: affected individuals and civil society actors were entirely excluded from the model’s design, implementation, and review phases.

This procedural issue reflects what Barocas, Hardt, and Narayanan (2019) refer to as the pitfalls of measurement: models trained on observational data cannot be assumed fair by default, especially when historical data reflect societal biases. The risk model used past enforcement decisions as training input—decisions that were themselves shaped by human biases and systemic inequality (Autoriteit Persoonsgegevens, 2020; Amnesty International, 2021). When asked about fairness audits or pre-deployment assessments, algorithm experts noted the absence of any procedural safeguards. A machine learning engineer involved in public sector contracting mentioned:

"There was no fairness testing. No bias audits. Not even basic demographic breakdowns of false positives. It was accuracy-obsessed—and not even good at that."

One of the journalists emphasized that the Dutch Tax Administration operated in a vacuum of ethical oversight:

"They treated this like an internal fraud detection tool. Fairness was conceptually missing."

The algorithm's deployment cannot be analyzed apart from the institutional context. As Busuioc and Alon-Barkat (2023) observe, public officials often experience automation bias—a tendency to over-trust algorithmic outputs even when flawed. In the Netherlands, civil servants were under pressure to increase fraud detection (Amnesty International, 2021, p. 13). Success was measured not in social protection, but in recovered funds (Amnesty International, 2021, p.12-14). This created what Amnesty International (2021, p. 13) calls a perverse incentive structure, where the algorithm became a revenue-recovery tool rather than a neutral risk filter.

Experts and journalists also stressed that fairness was undermined not only by the algorithm, but by the institutional logic surrounding it. One journalist noted:

"There was pressure to deliver fraud numbers. If the algorithm flagged more cases, that looked like success. Nobody asked whether it was fair—just whether it was effective at catching people."

Moreover, no organizational culture of fairness counterbalanced the algorithm's logic. There was no internal challenge mechanism, no whistleblower protections that functioned, and no active monitoring of racial impacts (Parliamentary Interrogation Committee on Childcare Allowance, 2020). As the Parliamentary Interrogation

Committee on Childcare Allowance (2020, pp. 75–77) concluded, “signals from within the Tax Authority about wrongful sanctions were ignored or suppressed, and the organizational culture discouraged dissent”. The result was a technocratic regime that prioritized efficiency, deterrence, and fiscal control over social justice—a policy rationality incompatible with fairness (Eubanks, 2018; Benjamin, 2019).

Not only were families wrongly flagged; they were also systematically denied any alternative—clarification requests by families were routinely ignored (Parliamentary Interrogation Committee on Childcare Allowance, 2020, pp. 80–81). The algorithm’s logic was treated as proprietary or opaque by tax authority officials and policymakers (Amnesty International, 2021, p. 13). Even lawyers and journalists faced obstacles in accessing documentation (Parliamentary Interrogation Committee on Childcare Allowance, 2020). This eroded both procedural fairness and the right of citizens to know why they are being targeted, and to contest the validity of that targeting (Barocas, Hardt & Narayanan, 2019).

This aligns with empirical findings in Starke et al. (2022) that perceptions of algorithmic unfairness are intensified when individuals are denied explanations, voice, and control. Fairness, then, is not only about inputs and outputs—but about the legitimacy of the process as experienced by those affected.

The Dutch childcare benefits algorithm was technically defective and it operated without grounding in the ethical, legal, or democratic principles that should structure decision-making in the public sector. By failing to incorporate fairness as a design principle, and by embedding it in a governance structure that privileged suspicion over support, the state enacted algorithmic injustice on a massive scale (Amnesty International, 2021, pp. 12–15). This case underscores that fairness in public decision-making should not be optional—it is the foundation of democratic legitimacy and social trust.

Future systems must go beyond harm mitigation and strive for positive fairness objectives, including the protection of marginalized groups, transparency in scoring criteria, and participatory governance. As Diakopoulos’ framework suggests, fairness is not a technical problem alone—it is a political commitment to justice embedded in sociotechnical systems.

5.2 Transparency

Transparency, as conceptualized by Diakopoulos (2016), is a foundational principle of algorithmic accountability that enables stakeholders—citizens, policymakers, journalists, and oversight bodies—to scrutinize, question, and understand algorithmically mediated decisions. It encompasses both the internal logic of algorithms (e.g., inputs, model structure, decision rules) and the institutional context in which they are deployed (e.g., who is responsible, under what policy mandate, with what degree of oversight). In the Dutch childcare benefits scandal, this principle was neglected, resulting in systemic opacity that insulated the algorithm from both public and administrative scrutiny (Parliamentary Interrogation Committee on Childcare Allowance, 2020).

The childcare benefits risk classification model operated as a black box system—one in which the inputs and outputs were observable (i.e., who was flagged, and when), but the decision logic remained inaccessible. Even civil servants tasked with enforcing the system lacked meaningful understanding of how risk scores were generated (Parliamentary Interrogation Committee on Childcare Allowance, 2020).

From a public governance perspective, this opacity is particularly concerning. Algorithms used in the public sector should, by definition, be subject to a higher standard of transparency, given their coercive power over citizens' lives. Yet, neither the model's architecture nor the feature weights used in scoring were ever disclosed—not to the public, not to oversight bodies, and not even to internal auditors during its initial years of operation (Autoriteit Persoonsgegevens, 2020).

This complete lack of transparency violates Diakopoulos' (2015) call for demand-driven and proactive transparency, in which governments provide information about automated systems before and when they affect people. Instead, Dutch authorities refused to reveal how the algorithm operated, even in response to formal parliamentary inquiries, Freedom of Information requests, and litigation. The Parliamentary Interrogation Committee (2020) reported that the use of the risk model was never disclosed to Parliament on time, despite repeated requests dating back to 2017 (Parliamentary Interrogation Committee on Childcare Allowance, 2020, pp. 90–92). Amnesty International documented that both parents and journalists were denied access to technical documentation, making it impossible to verify or contest the algorithm's outputs (Amnesty International, 2021, pp. 13–15). Similarly, the Dutch Data Protection Authority (Autoriteit Persoonsgegevens) confirmed that the Tax Authority delayed and selectively provided documentation during its formal investigation, thereby obstructing effective oversight (Autoriteit Persoonsgegevens, 2020, p. 29). These actions exemplify a deliberate opacity strategy that insulated the system from scrutiny—even as its decisions caused widespread harm.

The opacity surrounding the childcare benefits algorithm was not accidental—it was institutionally maintained and strategically protected. The Dutch Tax authorities repeatedly invoked data protection and fraud prevention rationales to avoid public disclosure (Autoriteit Persoonsgegevens, 2020). This behavior reflects what Levy, Chasalow, and Riley (2021) identify as a critical shortcoming in algorithmic governance: institutions design accountability gaps that make it unclear who is responsible for what, thereby shielding themselves from liability.

From as early as 2017, when complaints about wrongful accusations intensified, multiple government bodies—including the Ministry of Finance and the Dutch Data Protection Authority—were denied or delayed access to key documentation related to the childcare benefits risk classification model (Autoriteit Persoonsgegevens, 2020). Despite growing public concern and mounting political pressure, the Tax Authority refused to share details about the algorithm's design, feature weights, or decision rules, even during formal investigation. For example, in 2019–2020, the Data Protection Authority reported serious delays and non-cooperation from the Tax Authority during its inquiry into discriminatory data processing (Autoriteit Persoonsgegevens, 2020). Similarly, the Parliamentary Interrogation Committee found that relevant documents were not disclosed to Parliament or oversight bodies until late in the inquiry process—well after the scandal had become public (Parliamentary Interrogation Committee on Childcare Allowance, 2020, pp. 90–92). Amnesty International (2021, pp. 13–15) further observed that systemic opacity persisted into 2021, with civil society, journalists, and legal professionals still unable to obtain meaningful access to algorithmic documentation . As a result, public understanding of the algorithm's role in welfare enforcement remained minimal until the scandal broke open in 2018. By that time, over 26,000 families had been falsely accused of fraud, many with no knowledge of the tool that condemned them.

According to journalists interviewed, it took years of investigative work and parliamentary pressure to even confirm the system's core functionality.

"We filed FOIA requests, we asked in committee meetings—everything came back redacted or refused," said one journalist. "The tax office acted like the model didn't exist."

This institutional stonewalling directly contradicts Diakopoulos' principle (2015) that transparency should not only be technical but sociotechnical—enabling institutional answerability, public deliberation, and democratic oversight (Diakopoulos, 2015, p.3) .

Algorithm experts I spoke with were stating that technical transparency alone would have been insufficient—even if the model were published, most citizens or journalists would lack the tools to interpret it. What mattered, they argued, was procedural transparency: who developed the model, what objectives guided it, what variables were used, how error rates were evaluated, and what recourse mechanisms existed. None of this was shared.

"You can publish all the Python scripts you want," said one interviewed AI specialist. "If you don't also publish who built it, why it was built, and how people can challenge it—it's not transparent. It's codewashing."

The refusal to disclose was further compounded by the incentive structure behind the algorithm. As Amnesty International (2021, p 9) notes, the Dutch Tax Authorities were under political and fiscal pressure to detect fraud and recover funds. The success of the algorithm was measured not in accuracy or fairness, but in amounts reclaimed from accused families.

This created a perverse logic where secrecy was functional: public knowledge of the algorithm's mechanisms could have enabled appeals, challenged risk classifications, or exposed discriminatory patterns—all of which could undermine the system’s enforcement efficiency. In effect, transparency was seen as a threat to operational control, rather than a requirement of public legitimacy (Amnesty International, 2021, p. 13).

One of the most damaging consequences of this opacity was that citizens were unable to understand or contest the decisions made against them. They were not told why they were flagged, what risk factors were used, or what evidence would be required to restore their benefits (Amnesty International, 2021, p. 11). This lack of transparency made it impossible for individuals to claim their rights—a violation of both democratic norms and basic administrative justice (Parliamentary Interrogation Committee on Childcare Allowance, 2020). Transparency, in this sense, is a precondition for both explainability and remedy. Without knowing how a decision was made, individuals cannot meaningfully appeal it, journalists cannot investigate it, and oversight bodies cannot audit it.

Lack of transparency also extended within the bureaucracy itself. Civil servants tasked with enforcing benefit recovery procedures had little to no knowledge of how risk scores were produced (Amnesty International, 2021, p. 27).

As one journalist I interviewed mentioned the interview he conducted with internal whistleblowers:

“The people applying sanctions didn’t know why someone was flagged. They just saw a score and acted. Some were uncomfortable, but they were told, ‘Trust the system.’” (Journalist, personal communication with internal whistleblower, 2023)

One of the investigating journalists echoed this:

“It wasn’t just a black box to the public. It was a black box to the people inside the system.”

This diffusion of ignorance effectively protected the algorithm from internal critique, even as it caused immense social harm. The failure of transparency in the Dutch childcare benefits scandal was not just an oversight, it was a structural denial of public knowledge, institutional accountability, and democratic participation. As Amnesty International (2021) found, the risk classification model's workings were deliberately kept hidden from parents, auditors, and oversight bodies.

Rebuilding public trust requires more than post-hoc disclosures; it demands a transparency-by-design regime in which all public-sector algorithms are documented, registered, and open to scrutiny. Transparency is not simply about seeing the code—it is about exposing the power structures behind it and reclaiming them for the public good (Ananny & Crawford, 2018, pp. 976–978; Diakopoulos, 2015, p. 3).

5.3 Explainability

Explainability is the capacity of an algorithmic system to provide understandable reasons for its decisions to those affected by them. In Diakopoulos' (2019, pp. 18-20) terms, it is not enough that a system is transparent in its logic or structure; it must also be able to communicate meaningful justifications to relevant stakeholders—including decision-makers, oversight institutions, and most critically, affected citizens. Explainability is the bridge between the technical logic of an algorithm and the human demands for justification, recourse, and trust.

In the Dutch childcare benefits scandal, this bridge was entirely absent (Amnesty International, 2021). The algorithm functioned as a black box and as a justification vacuum - a condition in which neither citizens nor administrators received reasons for the decisions made, leaving both citizens and administrators unable to grasp why decisions were made, what evidence was used, or how to challenge those outcomes (Autoriteit Persoonsgegevens, 2020, pp. 24–25; Parliamentary Interrogation Committee on Childcare Allowance, 2020, pp. 42–45).

Tens of thousands of families were accused of fraud, had their benefits suspended, and were ordered to repay large sums—without being provided any explanation as to why (Amnesty International, 2021, p. 10). Affected parents were frequently met with administrative silence or vague references to "irregularities" when they sought clarification. Requests for documentation, reasoning, or criteria were routinely denied or ignored (Amnesty International, 2021, pp. 10-11).

Victims described their interactions with the tax office as marked by ambiguity, silence, and blame-shifting. Many received letters demanding the return of tens of thousands of euros, with no explanation for what had triggered the accusation of fraud. Requests for clarification yielded only generic responses—or none at all (Amnesty International, 2021, pp. 13–15; Parliamentary Interrogation Committee on Childcare Allowance, 2020, pp. 45–48). As the interviewed victim stated:

“They told me I had to pay back 35,000 euros,” the interviewee recalled. “I asked them, ‘Why?’ They said there were ‘discrepancies,’ but no one would say what. I begged them to tell me what I did wrong. They said it wasn’t their job to explain.”

This created a Kafkaesque situation: people were punished by a system whose logic they could not see and whose decisions they could not understand (Amnesty International, 2021, p. 12). From a public policy standpoint, such an arrangement violates the fundamental administrative principle of justified decision-making, which holds those public actions—especially coercive ones—must be explained in a way that citizens can reasonably interpret and, if necessary, contest (Parliamentary Interrogation Committee on Childcare Allowance, 2020).

The failure of explainability extended beyond the citizen-state interface. Civil servants themselves, particularly frontline caseworkers, were unable to understand how or why the algorithm had flagged certain cases (Amnesty International, 2021, p. 11). They were provided only with output scores and limited metadata, often stripped of any rational or feature attribution (Parliamentary Interrogation Committee on Childcare Allowance, 2020, pp. 45-46). As a result, officials became functionaries of automation, acting on opaque signals without a clear understanding of their basis (Parliamentary Interrogation Committee on Childcare Allowance, 2020, p. 55).

Civil servants interviewed by the journalists I talked conveyed similar frustration. According to one of them:

“Several caseworkers told me they received flagged cases with no context. No variables, no weights, no rationale. Just: this is a high-risk case. Proceed.”

This left them unable to make informed decisions or exercise discretion. One algorithm expert I interviewed described this as “algorithmic outsourcing of judgment”—a bureaucratic workflow where humans executed decisions without understanding them:

“If the humans in the loop can’t explain the tool’s logic, they’re not really in the loop. They’re just rubber-stamping a machine.”

This dynamic not only undermined the possibility of human oversight—it removed explainability at every institutional level, trying to make the system more efficient but less just. As one of the ADM experts explained during the interview, the lack of explainability in the childcare benefits algorithm led to two connected problems. First, it contributed to automation bias— phenomenon where people tend to trust algorithmic outputs over their own judgment, even when there is reason to doubt the system (Eubanks, 2018, p. 45; Diakopoulos, 2019, p. 96). In this case, civil servants were often unsure how risk scores were produced but still treated them as authoritative. Second, the system encouraged the mechanical application of recovery procedures, meaning that frontline workers followed rigid instructions and applied sanctions automatically, without considering individual circumstances. Civil servants became more like executors of automated decisions than independent evaluators (Amnesty International, 2021). This situation limited their ability to act based on judgment or compassion.

At a structural level, the algorithm’s opacity was exacerbated by its technical complexity and lack of design for interpretability (Amnesty International, 2021). The model incorporated self-learning elements via statistical or machine learning methods—without embedding any mechanisms for feature explanation or counterfactual reasoning (i.e., “what if” analysis) (Autoriteit Persoonsgegevens, 2020, pp. 25-26). As one Dutch ADM expert interviewed for this study noted, “These techniques are now standard in explainable AI (XAI) practice, including SHAP, LIME, or simpler rule-based models”.

This design choice shows a deeper problem: the algorithm was built to be efficient at predicting who might commit fraud, but not to explain its decisions in a way that people could understand or challenge. As Barocas, Hardt, and Narayanan (2023) argue, many ADM systems are trained to maximize statistical fit rather than social intelligibility—a trade-off that, in public administration, often leads to silent injustice. One algorithmic design expert explained:

"They could have built in explainability. It's not rocket science. Even with complex models, you can generate local explanations."

Explainability is not just a technical or legal concern, it is a question of the rights of people to understand the decisions that affect them. As philosopher Miranda Fricker (2007, pp 1-10) argues, individuals have a moral right to be treated as knowers and to participate in the reasoning processes that affect them. In the Dutch case, this right was structurally denied. Parents were stripped not only of benefits but of the capacity to understand, respond to, or resist the decisions taken against them.

This exclusion constitutes what Diakopoulos (2016) and later scholars such as Wenzelburger (2022) identify as a democratic failure of algorithmic governance: a breakdown in the justificatory functions of the state. When people cannot understand why they are punished, they cannot meaningfully engage with the law, advocate for themselves, or hold power accountable. As Diakopoulos (2016) argues, explainability should be judged based on its practical impact, not just its technical features. In other words, it's not enough to document how a system works internally—what matters is whether the people affected by the system can actually understand and respond to the decisions it produces. Especially in cases involving punitive consequences, such as the Dutch childcare benefits scandal, systems must be explainable in ways that support real accountability, regardless of whether a human formally approves the decision. Any system that triggers punitive action must be explainable—regardless of whether a human clicked the final button.

The childcare benefits algorithm was not merely unjust or opaque—it was inexplicable by design. Its architecture and implementation foreclosed the possibility of meaningful understanding for citizens and civil servants (Amnesty International, 2021, pp. 13, 27). In doing so, it severed the link between power and justification—a core principle of democratic administration (Diakopoulos, 2015, p. 3).

True explainability demands more than documentation or audits. It requires systems that are intelligible to non-experts, reason-giving that is responsive to human concerns, and institutional processes that enable dialogue between state and citizen. Without this, algorithms remain silent arbiters of fate—technically efficient, but normatively illegitimate (Diakopoulos, 2015, p. 3; Ananny & Crawford, 2018, pp. 976–977).

The analysis of the Dutch childcare benefits scandal clearly shows that the algorithm used by the Dutch Tax and Customs Administration failed to meet basic principles of fairness, transparency, and explainability, as described in Diakopoulos' Algorithmic Accountability Framework (Diakopoulos, 2016; Diakopoulos & Friedler, 2021). The system disproportionately targeted vulnerable groups, especially parents with a migration background or lower income levels. Although the algorithm did not use ethnicity as a direct input, it relied on proxy features like dual nationality, language, and certain postal codes—factors that ended up discriminating against ethnic minorities (Autoriteit Persoonsgegevens, 2020; Amnesty International, 2021). These targeting practices occurred without procedural safeguards, such as impact assessments or fairness audits, undermining both distributive and procedural fairness (Autoriteit Persoonsgegevens, 2020, p. 26).

The whole system was also deeply opaque, both technically and institutionally. Officials refused to share how it worked, even when questioned by Parliament or asked by oversight bodies (Parliamentary Interrogation Committee, 2020; Autoriteit Persoonsgegevens, 2020). One ADM expert interviewed for this thesis, who had worked on algorithmic systems in Armenia, noted:

"Across many public-sector projects I've seen, the people implementing the system rarely know what's under it. The model is delivered as a finished product, and staff are told to just trust it."

The lack of explainability meant that both citizens and civil servants were left in the dark. Families who were accused of fraud didn't know what they were being punished for, and frontline workers had no ability to question the system or make case-by-case decisions (Parliamentary Interrogation Committee on Childcare Allowance, 2020, pp. 44–46).

The interview data further confirmed these institutional and technological failures, revealing that the algorithm was within a bureaucratic environment that favored operational efficiency and fraud detection metrics over citizen rights and due process. All four ADM experts I interviewed emphasized that algorithmic systems always involve trade-offs. For example, increasing explainability and transparency often comes at the cost of reduced efficiency, and vice versa. While the field is evolving and future developments may allow us to balance both, the experts agreed that at present, achieving full transparency and maximum efficiency at the same time is very difficult. As one expert put it, "We build what we're asked to build—it's up to governments to decide what the system should prioritize." This highlights the important point that it is not just a technical decision, but a political and ethical one. Public institutions must be aware of these trade-offs and take responsibility for defining what matters most when deploying ADM systems.

Overall, this case provides a cautionary example of how the interaction of defective system design and dysfunctional governance practices can lead to profound breakdown of administrative justice. The next chapter will critically reflect on the applicability and limits of Diakopoulos' framework in capturing the full complexity of these failures and propose recommendations for improving algorithmic accountability in public sector contexts.

6. Discussion

This case study applied Diakopoulos' Algorithmic Accountability Framework to analyze how the Dutch childcare benefits algorithm failed to meet core principles of fairness, transparency, and explainability. Drawing from qualitative interviews and secondary sources, the findings highlight a deeply embedded structural and procedural failure that led to systemic injustice.

In terms of fairness, the algorithm was discriminatory by design. Though explicitly neutral in code, it indirectly targeted parents with dual nationality, foreign-sounding names, or lower socio-economic status. The victim I interviewed said that: "It felt like I was being punished for who I am, not what I did." The system failed both in terms of individual fairness (treating similar individuals differently based on irrelevant factors) and group fairness (systematically disadvantaging vulnerable groups). Experts and journalists interviewed consistently characterized the lack of fairness assessments and absence of safeguards as evidence of institutional neglect rather than mere technical error.

The transparency failures were equally profound. The algorithm operated as a black box, shielded by deliberate institutional opacity. Victims, lawyers, journalists, and even civil servants were unable to understand or access the decision-making logic. Interviews and reports (Amnesty International, 2021; Autoriteit Persoonsgegevens, 2020) revealed some patterns of bureaucratic evasion, where information requests were frequently stonewalled or denied, often justified with vague references to protecting the effectiveness of fraud detection. One journalist described it as "a perfect machine for self-protection."

Regarding explainability, both citizens and bureaucrats were effectively excluded from any reasoning process. Decisions were communicated without justification, preventing citizens from exercising their right to appeal. Civil servants interviewed by the journalists I spoke to admitted that they mechanically followed algorithmic outputs without understanding the rationale behind risk classifications. This created a justification vacuum, undermining procedural justice and institutional legitimacy.

Overall, these failures were not solely technical or institutional; rather, they reflected an interaction of flawed design choices, poor governance, and perverse bureaucratic incentives. The findings align closely with Diakopoulos' framework in diagnosing accountability gaps, yet the depth of institutional complicity and bureaucratic culture of silence suggest that algorithmic accountability must also be seen as an organizational and political challenge.

Applying Diakopoulos' Algorithmic Accountability Framework to the Dutch childcare benefits scandal proved highly valuable in diagnosing the key accountability failures. The framework's clear focus on fairness, transparency, and explainability provided a structured lens for examining the complex failures of both the algorithm and the surrounding governance systems. It allowed this research to systematically assess not only the technical properties of the algorithm but also the institutional context and human impacts.

The strength of the framework lies in its emphasis on accountability as a multidimensional concept that goes beyond technical audits to incorporate stakeholder engagement, responsibility, and governance structures. This case strongly confirmed Diakopoulos' assertion that algorithms must be designed and deployed with fairness and explainability as fundamental requirements—not optional add-ons. The study also reinforced his call for proactive and demand-driven transparency, which was conspicuously absent in the Dutch case.

However, empirical findings from this research and interview data also exposed limitations in the framework. While Diakopoulos' model effectively captured the systemic failures, it did not fully account for the human-algorithm interaction dynamics that were so critical to the Dutch case. As Busuioc and Alon-Barkat (2023) argue, the accountability failures are not only algorithmic but are amplified by organizational culture and cognitive biases among frontline civil servants, such as automation bias and selective adherence to algorithmic recommendations. The framework overlooks the role of administrative actors who act as intermediaries between automated outputs and citizen outcomes.

Similarly, while the framework addresses fairness primarily as a distributive issue, it does not fully engage with the critical sociological dimensions of structural racism, institutional discrimination, and power asymmetries that defined the experience of victims in this case. Starke et al. (2022) show that fairness perceptions in algorithmic decision-making are deeply shaped by individuals' trust in public institutions and their sense of procedural legitimacy. Through survey experiments conducted in Germany and the U.S., they found that people were more likely to perceive an algorithmic decision as fair when they trusted the institution using the system and believed the process was transparent and accountable—elements which Diakopoulos acknowledges but does not place at the center of his framework.

Finally, The Dutch case illustrates that, in practice, institutions may deliberately resist or undermine transparency mechanisms to protect operational efficiency or avoid public scrutiny. This highlights the need to extend the Diakopoulos' framework to incorporate political economy and incentive structures behind algorithmic deployment in government settings. Diakopoulos' framework provided a foundation for this analysis but would benefit from integration with adjacent theories like Lipsky's (1980) street-level bureaucracy, which highlights how frontline workers mediate policy under constraints; Romzek and Dubnick's (1987) typology of accountability, showing the interplay between legal, political, and professional pressures or Eubanks (2018), who examines how ADM systems can entrench social inequalities. These theories account for human factors, administrative behaviors, and structural inequalities in algorithmic governance.

The findings of this study suggest that algorithmic accountability in public administration cannot be fully understood without also examining the institutional and political environment in which ADM systems operate. As scholars like Bovens and Schillemans (2014) and Dubnick (2005) argue, accountability is not merely a matter of information disclosure or outcome fairness, but is shaped by organizational incentives, bureaucratic routines, and the presence (or absence) of effective accountability forums. In the Dutch childcare case, the opacity of the algorithm was sustained not just by technical complexity, but by deliberate institutional strategies to limit disclosure, avoid blame, and maintain control over fraud detection narratives. These dynamics reflect what public administration literature identifies as symbolic accountability and blame avoidance—concepts discussed by Bovens and Schillemans (2014, pp. 6–7), who describe symbolic accountability as procedures that provide a façade of transparency without real scrutiny, and by Dubnick (2005, p. 382), who explores how accountability mechanisms can be co-opted to deflect blame rather than enhance responsibility. These phenomenas are not captured fully by Diakopoulos' principles alone. Moreover, the failure of internal oversight mechanisms and the marginalization of dissenting voices suggest that even well-designed accountability frameworks may falter in institutional environments that prioritize compliance, efficiency, or political protection over ethical governance. Thus, while Diakopoulos' model remains essential for diagnosing algorithmic harms, it should be complemented by public administration theories such as Lipsky's (1980) street-level bureaucracy, which highlights how frontline workers operate under constraints that often limit discretionary decision-making, Hood's (2010) work on administrative transparency, which questions the political motivations behind disclosure practices, and Romzek and Dubnick's (1987) typology of accountability relationships, which emphasizes the tension between hierarchical control, professional norms, and legal accountability. These frameworks help situate algorithmic systems within broader bureaucratic and political dynamics that shape their design, deployment, and oversight.

The Dutch childcare benefits scandal shows how dangerous algorithmic decision-making can be when it's not properly managed—and offers important lessons for governments and public administrators. This research demonstrates that technical sophistication alone cannot prevent harm if ADM systems are embedded within institutional logics that prioritize efficiency and control over fairness and justice.

First, governments must recognize that algorithmic systems in the public sector are not neutral tools. As this case illustrates, even semi-automated decision-support systems can result in disproportionate harm when deployed without robust safeguards. Policy responses must therefore move beyond technical risk assessments to include institutional accountability, participatory design, and oversight mechanisms from the very beginning of ADM system development.

Second, this case underscores the critical need for mandatory algorithmic impact assessments prior to deployment, especially in high-stakes welfare and social protection domains. These assessments must explicitly consider distributive impacts, potential biases, and harms to marginalized groups. Governments should also enforce the creation of public registries of ADM systems, listing their purpose, decision logic, and risk assessments, to promote transparency and public scrutiny (AlgorithmWatch, 2020, pp. 22–23).

Third, governments should institutionalize human oversight and discretion, ensuring that ADM systems cannot be used as unchallengeable instruments of enforcement. As interviewees in this study emphasized, the absence of an effective human appeal mechanism was as damaging as the flaws in the algorithm itself.

Fourth, the scandal illustrates the crucial role of civil society, investigative journalism, and whistleblowers in holding state algorithms accountable. Journalists and victims played a pivotal role in exposing the failings of the Dutch system.

Fifth, regulatory frameworks at both national and European levels must move towards a model of continuous accountability. As several algorithm experts in this study argued, static evaluations or pre-deployment audits are not enough. ADM systems must be subject to ongoing monitoring, post-deployment audits, and responsive governance structures that allow for system redesign if harms are detected.

Lastly, this case highlights that algorithmic accountability must be seen not only as a technical or legal issue, but as a core requirement of democratic governance. Public trust in digital government will depend on ensuring that automated systems remain understandable, contestable, and responsive to the people they serve.

6.2 Contributions and Limitations

This study makes a meaningful contribution to the growing scholarly discourse on algorithmic decision-making, digital governance, and public administration by providing a comprehensive, empirically grounded analysis of one of the most consequential algorithmic governance failures in Europe. It does so by applying and critically evaluating Diakopoulos' Algorithmic Accountability Framework, and by integrating qualitative data from multiple stakeholder perspectives—including victims, journalists, and algorithm experts.

The research extends the existing literature in several ways. First, it reinforces prior work by Barocas, Hardt, and Narayanan (2019) on how ADM systems, particularly when trained on biased historical data, can reproduce, and entrench existing patterns of inequality and discrimination. The findings from this case illustrate that biased data, combined with absent fairness safeguards and flawed bureaucratic incentives, can lead to systemic harm.

Second, this case study provides empirical support for recent work by Busuioc and Alon-Barkat (2023), who highlight the dangers of automation bias and selective adherence in semi-automated public systems. The Dutch childcare algorithm became a textbook example of how civil servants over-trusted opaque systems and systematically removed human discretion, even when faced with evidence of harm.

Third, it complements the literature on fairness perceptions in ADM, particularly as articulated by Starke et al. (2022). This study shows that lack of procedural transparency and the denial of explanation rights amplified perceptions of unfairness, not only among affected parents but also among civil society actors and frontline bureaucrats.

Fourth, the research enriches the public administration literature by demonstrating that ADM failures cannot be fully understood without accounting for organizational cultures, institutional incentives, and political contexts. It suggests that traditional public administration frameworks, such as those of Bovens (2007), must be adapted to better capture the unique accountability challenges posed by ADM systems, which often blur lines of responsibility and hinder effective oversight.

Finally, this thesis responds to calls within AI ethics and governance research for more in-depth, contextualized case studies of real-world ADM deployments. By providing a detailed analysis of the Dutch childcare benefits scandal, this research contributes to the empirical literature by applying Diakopoulos's accountability principles to examine how the absence of fairness, transparency, and explainability in ADM can lead to harmful outcomes. It further explores why these principles were not upheld by drawing on the broader concept of algorithmic governance. In doing so, the study provides actionable insights for scholars, policymakers, and practitioners working at the intersection of digital government and democratic accountability.

While this study offers an in-depth and multi-perspective analysis of the Dutch childcare benefits scandal, it is important to acknowledge its limitations. First, as a single case study, the findings may not be fully generalizable to all ADM systems or public sector contexts. The Dutch case was shaped by specific institutional, legal, and cultural factors that may differ from other countries' administrative environments.

Second, the study relies heavily on qualitative methods, including primary interviews and secondary document analysis. Although triangulation was applied to enhance validity, qualitative interviews carry the risk of selection bias and subjectivity in interpretation. The study does not include interviews with system designers or high-level officials, which may have constrained insights into internal decision-making processes.

Third, while Diakopoulos' framework provided a valuable structure for analysis, its focus on three core principles (fairness, transparency, explainability) may have limited the study's ability to fully explore broader systemic and political economy factors, such as regulatory capture, public-sector austerity pressures, and structural racism as highlighted by scholars like Benjamin (2019) and Eubanks (2018).

Future research should address these limitations in several ways. Comparative studies could be conducted across different countries to explore whether similar patterns of ADM accountability failure occur under varying institutional and cultural conditions. Such comparative research could also examine whether countries with more robust regulatory frameworks (e.g., those under the EU AI Act) experience fewer or different types of harm.

Additionally, there is a need for more interdisciplinary work that combines technical AI auditing methods with institutional analysis from public administration and political science. Future research could also focus on developing and empirically testing expanded accountability frameworks that incorporate organizational dynamics, media influence, and civil society engagement as part of the algorithmic governance ecosystem.

Finally, longitudinal studies examining the long-term effects of algorithmic harms on public trust, institutional legitimacy, and citizen-state relations would provide valuable insights into the ongoing social costs of ADM failures. Despite these limitations, this thesis provides a strong foundation for further inquiry into algorithmic accountability and offers actionable insights for policymakers and scholars seeking to prevent similar governance failures in future deployments of ADM technologies.

7. Conclusion

This thesis set out to critically examine the Dutch childcare benefits scandal as a landmark case of accountability failure in algorithmic decision-making within public administration. The primary research question explored how and to what extent the Dutch childcare benefits algorithm failed to meet the principles of fairness, transparency, and explainability as proposed by Diakopoulos' Algorithmic Accountability Framework. Using a qualitative case study approach that combined documentary analysis and qualitative interviews with victims, journalists, and experts, the research provides new empirical insights into the scandal.

The findings revealed that the childcare benefits algorithm violated all three core principles of accountability. In terms of fairness, the system systematically discriminated against vulnerable groups, particularly parents with dual nationality or ethnic minority backgrounds, without any meaningful safeguard mechanisms. The transparency failures were characterized by extreme opacity at every institutional level; not only were citizens denied information about why they were targeted, but even civil servants lacked the ability to understand or explain the outputs of the system. The absence of explainability created a justification vacuum, leaving citizens powerless to understand, contest, or appeal decisions that had devastating financial, social, and psychological impacts.

The research confirmed the value of Diakopoulos' framework as a diagnostic tool. It provided a structured way to analyze both the technical and governance failures that contributed to the scandal. However, the study also identified important limitations: the framework does not fully account for the organizational and institutional dynamics—such as automation bias, perverse incentives, and bureaucratic self-protection—that were so critical to understanding how the harms were amplified in practice. This study reinforces and extends existing critiques of Diakopoulos' framework by showing, through empirical analysis and theoretical integration, that algorithmic accountability must better account for human-algorithm interaction dynamics and the institutional power asymmetries embedded in ADM systems (Busuioc & Alon-Barkat, 2023; Eubanks, 2018; Ananny & Crawford, 2018)

The Dutch case offers vital lessons for policymakers and public institutions. First, ADM systems must never be implemented in high-stakes contexts such as welfare enforcement without robust pre-deployment impact assessments and ongoing audits. Second, transparency must be proactive and meaningful, with governments establishing public ADM registries and clear redress mechanisms for citizens. Third, human oversight cannot be replaced by automated systems; discretion, empathy, and procedural justice must remain cornerstones of public decision-making. Finally, this research underscores the essential role of civil society, media, and investigative journalism as accountability actors when formal institutional safeguards fail.

This study contributes to the growing literature at the intersection of public administration, digital governance, and AI ethics by demonstrating how Diakopoulos' Algorithmic Accountability Framework can be applied empirically to analyze accountability failures in a real-world public sector context. It demonstrates that algorithmic accountability is not only a technical or legal issue, but a profoundly political and institutional challenge. The Dutch childcare benefits scandal serves as a critical warning of how public-sector ADM systems can produce catastrophic social harms when ethical design, regulatory oversight, and participatory governance are absent.

The research also identifies pathways for further study. Comparative case studies of ADM governance in other national contexts could help validate and refine accountability frameworks. Future research should also explore how regulatory regimes like the EU AI Act can operationalize continuous accountability in the deployment of ADM systems. Moreover, interdisciplinary approaches that integrate public administration theory with machine learning audit methodologies may provide richer tools for both scholars and practitioners.

In conclusion, the Dutch childcare benefits scandal should be remembered as a failure of governance and as a failure of an algorithm. It stands as a cautionary tale that highlights the ethical responsibility of public institutions to ensure that emerging technologies serve the public interest and uphold the fundamental principles of fairness, transparency, participation, and human dignity in democratic societies.

References:

AlgorithmWatch. (2020). Automating Society Report 2020.

Amnesty International. (2021). Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal. Amnesty International Netherlands.

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

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against Blacks. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Autoriteit Persoonsgegevens. (2020). Onderzoek Belastingdienst/Toeslagen: De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. https://fairmlbook.org/

Benjamin, R. (2019). Race after technology: Abolitionist tools for the New Jim Code. Polity Press.

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. European Law Journal, 13(4), 447–468.

Bovens, M., & Schillemans, T. (2014). Meaningful accountability. In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), The Oxford Handbook of Public Accountability (pp. 1–20). Oxford University Press.

Diakopoulos, N. (2015). Algorithmic accountability reporting: On the investigation of black boxes. Tow Center for Digital Journalism.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. Communications of the ACM, 59(2), 56–62. https://doi.org/10.1145/2844110

Diakopoulos, N. (2019). Automating the news: How algorithms are rewriting the media. Harvard University Press.

Diakopoulos, N., & Friedler, S. A. (2021). How to hold algorithms accountable. MIT Sloan Management Review, 62(2), 1–5.

Denzin, N. K., & Lincoln, Y. S. (2011). The SAGE handbook of qualitative research (4th ed.). Sage Publications.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Dubnick, M. J. (2005). Accountability and the promise of performance: In search of the mechanisms. Public Performance & Management Review, 28(3), 376–417.

ECNL. (2022). Netherlands sets precedent for human rights safeguards in use of AI. European Center for Not-for-Profit Law. https://ecnl.org/news/netherlands-sets-precedent-human-rights-safeguards-use-ai

European Commission. (2019). Standard Eurobarometer 91: Public opinion in the European Union – Spring 2019. https://europa.eu/eurobarometer/surveys/detail/2255

European Union. (2016). General Data Protection Regulation, Article 22: Automated individual decision-making, including profiling.

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Fenster, M., Ananny, M., & Diakopoulos, N. (2020). Transparency. In The Oxford Handbook of Ethics of AI. Oxford University Press.

Fuster, A., Plosser, M., Schnabl, P., & Vickery, J. (2022). Predictably Unequal? The Effects of Machine Learning on Credit Markets. Journal of Finance, 77(1), 5–47.

Hood, C. (2010). Accountability and transparency: Siamese twins, matching parts, awkward couple? West European Politics, 33(5), 989–1009.

Konaté, S., & Pali, B. (2023). "You have to talk with us, not about us": Exploring the harms of wrongful accusation and possibilities for transformative justice in the Dutch childcare-benefits scandal. Revista de Victimología / Journal of Victimology, 16, 139–164.
https://pure.uva.nl/ws/files/136071877/276_767_1_PB.pdf

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165(3), 633–705.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. ProPublica.
https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36–43.

Lipsky, M. (1980). Street-level bureaucracy: Dilemmas of the individual in public services. Russell Sage Foundation.

Margetts, H., & Dunleavy, P. (2013). The second wave of digital-era governance: A quasi-paradigm for government on the web. Philosophical Transactions of the Royal Society A, 371(1987).

OECD. (2019). OECD Digital Government Review of Sweden. OECD Publishing.

OECD. (2021). The OECD AI Policy Observatory.

Oswald, M., Grace, J., Urquhart, L., & Hutton, W. (2018). Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and 'Experimental' Proportionality. Information & Communications Technology Law, 27(2), 223–250.

Oxford University Press. (n.d.). AI. In Oxford Learner's Dictionaries. Retrieved March 9, 2025, from https://www.oxfordlearnersdictionaries.com/definition/english/ai

Oxford University Press. (n.d.). Machine learning. In Oxford Learner's Dictionaries. Retrieved March 9, 2025, from https://www.oxfordlearnersdictionaries.com/definition/english/machine-learning

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Publishing.

Parliamentary Interrogation Committee on Childcare Allowance. (2020). Ongekend onrecht: Verslag parlementaire ondervragingscommissie Kinderopvangtoeslag. Tweede Kamer der Staten-Generaal.

Pasquale, F. (2015). The black box society: The secret algorithms that control money and information. Harvard University Press.

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency, 469–481.

Richardson, R., Schultz, J., & Crawford, K. (2019). Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. NYU Law Review, 94, 192–233.

Romzek, B. S., & Dubnick, M. J. (1987). Accountability in the public sector: Lessons from the Challenger tragedy. Public Administration Review, 47(3), 227–238.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Schillemans, T. (2011). Does horizontal accountability work? Evaluating potential remedies for the accountability deficit of agencies. Administration & Society, 43(4), 387–416.

Snellen, I. (2005). ICTs, bureaucracies, and the future of democratic governance. In V. Bekkers, H. van Duivenboden, & M. Lips (Eds.), Governance and the democratic deficit. Routledge.

Starke, C. et al. (2022). Fairness Perceptions of Algorithmic Decision-Making. Big Data & Society.

Veale, M., & Brass, I. (2019). Administration by Algorithm? Public Management Meets Public Sector Machine Learning. Public Administration Review, 79(6), 811–825.

Weller, A. (2019). Transparency: Motivations and challenges. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning.

Yanow, D., & Schwartz-Shea, P. (2014). Interpretive research design: Concepts and processes. Routledge.

Yeung, K. (2018). Algorithmic regulation: A critical interrogation. Regulation & Governance, 12(4), 505–523.

Zarsky, T. Z. (2016). The trouble with algorithmic decisions: An analytic roadmap to examine efficiency and fairness in automated and opaque decision-making. Science, Technology, & Human Values, 41(1), 118–132.

Ziewitz, M. (2016). Governing algorithms: Myth, mess, and methods. Science, Technology, & Human Values, 41(1), 3–16.

* During the preparation of this thesis, the author used the AI tool *ChatGPT* developed by OpenAI in order to support text formulation, structure clarification, and reference formatting. All outputs generated by the tool were critically reviewed, edited, and verified by the author, who takes full responsibility for the accuracy and academic integrity of the final content.

List of Appendices

Appendix 1: Interview Protocol

Interviewer: Tigran Karapetyan
Participant ID: [To be assigned: Participant A, B, C, etc.]
Date/Time:
Interview Medium: Zoom (online video conferencing)

Interview Introduction Script

Thank you for agreeing to participate in this interview. My name is Tigran Karapetyan, and I am conducting this research as part of my thesis at Charles University to explore accountability issues related to algorithmic decision-making, particularly the Dutch childcare benefits scandal. The purpose of this interview is to understand your perspective and experiences related to fairness, transparency, and explainability of the ADM system used in this case.

This interview will take approximately 30 minutes. Your responses will remain confidential, and your identity will be anonymous in the thesis and any related publications. You have the right to skip any question, and you may stop the interview at any time.

Do you have any questions before we start?

Part 1: Participant Background (5 minutes)

1. Can you briefly describe your role or involvement with the Dutch childcare benefits case?
2. How long have you been engaged with or affected by this issue?

Part 2: Explainability (5–10 minutes)

3. From your perspective, how clearly were algorithmic decisions explained to stakeholders involved (victims, public officials, policymakers)?
    - Could you or other stakeholders easily understand how and why certain individuals were flagged by the system?

    - Were adequate explanations provided when stakeholders sought clarification about algorithmic outcomes?
4. Can you recall any specific incidents or examples where explanations about the ADM system were particularly unclear, inadequate, or helpful?
5. In your view, how did the lack of clear explanations affect stakeholders' ability to challenge or correct potential errors or unfair outcomes?

Part 3: Transparency (5–10 minutes)

6. How transparent was the Dutch government regarding the childcare benefits ADM system (e.g., criteria, data sources, procedures)?
    - Did you or other stakeholders have access to sufficient information about how the ADM system operated?
    - Were details of the ADM system publicly available, or was the information difficult to access?
7. Did decision-makers or oversight bodies have sufficient transparency to perform their accountability roles effectively? Why or why not?
8. What factors do you think contributed to the transparency gaps, if any, regarding the ADM system?

Part 4: Fairness (5–10 minutes)

9. Based on your experience, did the ADM system unfairly target or disadvantage certain groups or individuals?
    - If yes, which demographic, ethnic, or socio-economic groups were most affected?
    - Do you think these unfair outcomes were predictable or preventable?
10. Were there mechanisms or safeguards in place to identify, mitigate, or prevent unfair or discriminatory outcomes from occurring?
    - If yes, were these mechanisms effective? Why or why not?
    - If not, what types of safeguards do you believe should have been implemented?
11. In your opinion, what was the impact of unfair algorithmic decisions on affected individuals and communities?

Part 5: Reflections and Recommendations (5–10 minutes)

12. Reflecting on this experience, what would you say are the most critical lessons about fairness, transparency, and explainability in ADM that policymakers and implementers should learn from this case?
13. Are there any specific changes or recommendations you would suggest to improve accountability in the future use of ADM systems in the public sector?
14. Is there anything additional about your experience or insights related to this topic you would like to share?

Interview Closing Script

Thank you very much for your valuable time and insights. Your experiences and perspectives will contribute significantly to understanding and improving accountability in algorithmic decision-making. If you have further thoughts later or if you have additional contacts to recommend for this study, please feel free to contact me. Again, all your information will remain confidential and anonymous. Thank you for your cooperation.

Appendix 2: Informed Consent Form

Title of Research Project:
Accountability in Algorithmic Decision Making: Case Study Analysis

Researcher:
Tigran Karapetyan
Charles University, Institute of Sociological Studies, Department of Public and Social Policy

Supervisor:
Mirna Jusić, M.A., Ph.D.

Purpose of the Research
The thesis aims to conduct a case study analysis of accountability mechanisms in the Dutch childcare benefits scandal. It will evaluate how effectively the Dutch childcare benefits algorithm adhered to Diakopoulos' principles of Fairness, Transparency, and Explainability. Furthermore, the thesis will propose actionable strategies to better integrate these algorithmic accountability principles into public-sector decision-making systems.

Your Participation
You are being invited to participate in a semi-structured online interview. The interview will take approximately 30 minutes and will be conducted remotely via Zoom.

The interview will cover topics including:

- Your experiences and understanding related to the explainability of ADM decisions.
- The transparency of the ADM system used in the Dutch childcare benefits case.
- Perceptions and experiences regarding fairness and potential discriminatory outcomes of the ADM system.

Voluntary Participation and Right to Withdraw
Your participation in this research is entirely voluntary. You have the right to:

- Decline to answer any question(s).
- Pause or terminate the interview at any time.
- Withdraw your consent and have your data excluded at any point before the final submission of the research thesis.

Confidentiality and Anonymity
All information you provide will remain strictly **confidential**. To protect your identity:

- No identifying personal information will be included in the final thesis or any publications.

- You will be identified in transcripts and reports only by a code (e.g., "Participant A," "Participant B") along with a generalized description of your professional role (e.g., "journalist," "public official," "lawyer," etc.).
- Interview recordings and transcripts will be securely stored and accessible only to the researcher and, if necessary, their supervisor.

Contact Information

Should you have any further questions, concerns, or complaints regarding this study or your participation, please contact:

Researcher:

Tigran Karapetyan

Participant Agreement

By signing below, I confirm that:

- I have read and understood this Informed Consent Form and had the opportunity to ask questions, which have been answered satisfactorily.
- I voluntarily agree to participate in this research.
- I consent to being recorded for the purposes outlined above.
- I understand that I can withdraw my consent at any point during or after the interview, up until the final submission of the thesis.

Participant's Name: _______________________________

Participant's Signature (Digital or Typed Accepted): _______________________________

Date: ____________________________

Thank you very much for your willingness to participate in this research.