Vrije Universiteit Amsterdam

Bachelor Thesis

Concealing algorithms, biased data and unethical risk models -- the case of the Dutch Child Welfare Scandal

Author: Asiea Alrefai (2660652)

1st supervisor:	Dr. Anna Bon
daily supervisor:	Dr. Anna Bon
2nd reader:	Prof. Dr. Hans Akkermans

A thesis submitted in fulfillment of the requirements for the VU Bachelor of Science degree in Computer Science

July 28, 2023

Table of content:

Table of content: 1

Abstract: 3

Chapter 1: The use of AI and Automated Decision-Making in Public Administration: 4

1.1 AI in Public Administration: 4

1.2 Approach and Methodology: 5

Chapter 2: The Case of the Dutch Child Welfare Scandal 6

2.1 A Dutch Nation-wide Scandal 6

Chapter 3: About Black Box - Machine Learning Algorithms 8

3.1 Black Box Algorithms and How they are Used 8

3.2 Black Box Algorithm Development Stages 9

3.3 Relationship between Bias, Precision, and Accuracy and Black Box Algorithms 10

3.4 Conclusion: Importance of Cautious use of AI in Decision-making 13

Chapter 4: Conceptually Modeling the Workflow and the Stakeholders 14

4.1 About Activity Diagrams 14

4.2 Stakeholders Analysis 16

4.3 Analyzing the TVS workflow 18

4.4 Black Box AI model in TVS 20

4.4 Limitation of the UML Model 21

Chapter 5: Unpacking the System and Identifying Critical Flaws 22

5.1 The Role of the Black Box AI System in TVS 22

5.2 The SyRI System: using Big Data for Human Profiling 22

5.3.1 The indicators used in the Risk Classification Model: 25

5.3.1.1 The indicator Dutch nationality Yes/No: 26

5.3.1.2 The Age BSN indicator: 28

5.3.1.3 The FVS indicator: 29

5.3.1.4 The Income indicator: 29

5.3.1.5 Single parent and amount of hours in childcare indicators: 30

5.3.2 Issues of Population and Sample Size for the Risk Model: 31

5.4 Conclusion of this Chapter: Lack of Control Mechanisms 33

Chapter 6: Governmental Challenges in TVS Model Management and Transparency: 34

6.1 Unfulfilled Stakeholder Goals 34
6.2 Transparency and Communication Challenges 34
6.3 Concealment of the Algorithmic Process 35
6.4 Conclusion of this Chapter: Lack of Transparency 36
Chapter 7: Consequences and Impact 37
7.1 Output of the Risk Classification Model 37
7.2 The Chosen Indicators Leading to Discriminatory Results: 39
7.3 Consequences of the Dutch Welfare Scandal for Citizens: 40
7.4 Judicial Decision Prohibiting SyRI 42
7.5 Conclusion of this Chapter: Discriminatory Outcomes Affecting Many People 43
Chapter 8: Human Empathy and Automated Decision Making: 44
8.1 Diminished Sense of Responsibility: 44
8.2 Impact on Empathy and Human Connection: 45
8.3 Perceived Technological Determinism: 45
8.4 Liability Concerns: 45
8.5 Conclusion of this Chapter: Fostering Empathy and Responsibility in the Welfare Benefit System 46
Chapter 9: Discussion and Recommendations 47
9.1 Risks of Black Box Systems and Self-Learning Algorithms in Public Sector: 47
9.2 Modeling Techniques for the Sake of Openness: 47
9.3 Safeguarding Algorithmic Decision-Making 48
9.4 Strengthening Human Rights Oversight: 49
9.5 Ethnicity and Nationality in Law Enforcement Risk Profiling: 49
9.6 Seeking Justice in Algorithmic Systems: 50
9.7 Mitigate Bias in Black Box Algorithms: 50
9.8 Need for Transparency 52
9.9 Achieving Adequate Representation in Training Data 53
9.10 Need to Comply to Data Protection Rules (GDPR) 54
9.11 Need for Procedures of Ethical Clearance 55
9.12 Conclusion of this Chapter: Concluding Recommendations 56
Chapter 10: Conclusion 57
References: 59
Appendix: 66

Abstract:

[*Problem statement]* There are growing concerns about the use of AI algorithms by public authorities. Despite the rapid introduction and widespread use of AI in government agencies, no proper risk assessment is carried out prior to deployment. One of the main problems with the use of AI is the lack of transparency of machine learning models in the decision-making process. This obscures the arguments for scrutiny of fair, impartial and transparent decision-making and blurs responsibilities. This presents a serious risk in a democratic society.

[*Objective*] This research aims to understand how algorithmic decision-making system is currently being, used in public administration, and what are the risks, when this is applied to decision-making that affects human lives.

[*Context*] Therefore, in this thesis we evaluate the case study of the Dutch child welfare scandal. In this case, which took place between 2004 and 2019, the Dutch government's tax and customs administration used AI algorithms to assess citizens' potential fraud. As a result of a biased AI system, around 26,000 Dutch citizens were wrongly accused of fraud. The consequences of these accusations have plunged many citizens and their children into serious social, emotional and financial hardship for many years.

[*Methods*] To do so, workflows and decision-making processes are analyzed and reverse-engineered, through conceptual modeling. The conceptual models clarify the whole process and make it possible to detect biases and unethical decision-making.

[*Contribution*] This research project contributes to a better understanding of the potential risks in the use of AI in public administration. It provides a method to clarify complex work processes through the use of conceptual modeling techniques, which are common in computer science. It gives recommendations, how to make AI-supported decision-making in public administration fairer and more transparent.

[*Impact*] The case of the Dutch child protection services is a cautionary tale that highlights the importance of making AI systems explainable, fair and accountable. To address these concerns, it is imperative that policymakers, computer scientists and child protection experts collaborate and design algorithms with strong safeguards against discrimination and bias, and promote transparency, accountability and human rights in AI-based decision-making processes. This research gives some directions how this can be done

Chapter 1: The use of AI and Automated Decision-Making in Public Administration:

In this chapter, the utilization of AI and automated decisions in public administration will be discussed, along with an exploration of both their benefits and drawbacks. Subsequently, section 1.2 will provide a comprehensive elaboration and explanation of the approach and methodology employed in the research.

1.1 AI in Public Administration:

The integration of Artificial Intelligence (AI) and Automated Decision-Making (ADM) in public administration has witnessed substantial growth in recent years. This adoption has been driven by the potential benefits of AI in terms of efficiency and decision-making through machine learning techniques [36]. ADM involves the use of algorithms and AI to support or replace human decision-making in various areas of the public sector [37]. Semi-automated systems, which combine human judgment with automation, are also employed to arrive at decisions [38].

The widespread adoption of AI and ADM in the public sector is attributed to the rapid development and availability of advanced technologies and the increasing volume of data collected by government agencies. Policymakers view these technologies as promising solutions that offer effectiveness, efficiency, and cost-effectiveness. Additionally, algorithms are perceived as neutral decision-makers, free from human biases and limitations, which has fueled their use in critical areas like law enforcement and criminal justice [39].

However, The integration of AI and AMD in public administration presents both opportunities and challenges. One major concern is algorithmic bias, where AI systems learn from historical data and perpetuate existing biases, leading to discriminatory outcomes and the potential automation of inequality [39].

Additionally, the opacity of AI decision-making mechanisms poses accountability challenges. The "black box" nature of algorithms makes it difficult to ascertain who is ultimately responsible for specific decisions, especially when AI-driven outcomes have significant implications for individuals or communities.

Addressing these issues is crucial as they can have severe consequences, particularly for vulnerable and disadvantaged citizens, leading to discriminatory treatment and eroding trust in government. The potential for biased information processing and the selective adoption of algorithmic advice based on pre-existing stereotypes further exacerbates these concerns.

An illustrative example of the consequences of incautious ADM use is the Netherlands child welfare scandal, where an automated system led to unjustified decisions affecting families, highlighting the need for caution and accountability when employing AI in public administration.

1.2 Approach and Methodology:

The primary objective of this research is to study how artificial intelligence (AI) is utilized by the government, what the inherent risks are of applying ICT and AI in public administration, and which consequences this can have for citizens.

We do this by studying the case of the Dutch child welfare system, from which it is reported that discriminatory decisions have been produced by AI [2,3,4,6,9,12,21].

As the methodology, a combination of design science, literature study and ethnographic techniques (interviewing) were used. The approach was as follows.

First, we conducted an extensive literature review, encompassing online reports, investigations carried out by the government, parliamentary letters, articles, and research papers.

Secondly, an interview was conducted with an individual who had personally experienced the effects of the child welfare algorithm.

Thirdly, conceptual modeling was employed to provide a clear and illustrative representation of the complete process of the Dutch child welfare system. An activity diagram was produced after stakeholder analysis for which roles and goals of the various stakeholder groups have been extracted, based on the literature.

Using the produced conceptual models, the decision-making process was analyzed for the complete workflow; the AI and human-based decisions were assessed.

Next, the AI and ICT-based models were further unpacked and further assessed for possible bias, again, based on existing literature.

Finally, the influence of AI-based decision-making process on the stakeholders was discussed.

This is discussed in the following chapters.

Chapter 2: The Case of the Dutch Child Welfare Scandal

In this Chapter, we will delve into an extensive analysis of the Dutch child welfare scandal, examining it with a meticulous focus on its intricate details. The Dutch child welfare scandal serves as a poignant case study that sheds light on the intersection of algorithmic decision-making, government policies, administrative practices, and their real-world implications for individuals and families.

2.1 A Dutch Nation-wide Scandal

“I knew, it is just a big mistake that will get resolved once I call the Belastingdienst." Participant X's reaction when they received a letter from the Dutch tax authorities in June 2020, asking them to pay 8000 Euros because, as the tax authorities claim, they were not entitled to child welfare for the last year and a half.

The Dutch tax authorities introduced a childcare benefits program in 2005, granting eligibility to individuals residing or working in the Netherlands with children below 18 years of age. The purpose of this program is to provide financial support for the expenses associated with raising and caring for children. The government's contribution is determined based on the income level of the parents or caregivers, with higher levels of child benefits allocated to those with lower incomes [1]. In 2013, the tax authorities adopted an algorithm-based decision-making system called the risk classification model to detect and prevent fraudulent activities. This algorithmic system employed self-learning mechanisms to create risk profiles of childcare benefits applicants who were deemed more likely to submit inaccurate applications or engage in fraudulent behavior. When parents and caregivers were identified as potentially fraudulent by the system, their benefits were suspended, and investigations were initiated [4,6,9].

The tax authorities demonstrated the effectiveness of the algorithmic decision-making system by successfully recovering sufficient funds from alleged fraudsters to cover the costs of the operation. Consequently, the focus was on seizing funds to the maximum extent possible, irrespective of the accuracy of fraud allegations. To validate their eligibility for benefits, parents and caregivers were requested to provide additional evidence. However, when they sought clarification regarding the specific information considered incorrect or false, or the evidence that was missing, they often encountered a lack of transparency, as the tax authorities refused to elucidate their decision-making process. During this period, crucial information about the existence and functioning of the risk classification model remained inaccessible to various stakeholders, including parents, caregivers, journalists, politicians, and oversight bodies [9].

The Dutch tax authorities' implementation of the risk classification model resulted in a nationwide scandal, wherein a substantial number of parents and caregivers were erroneously implicated in cases of childcare benefit fraud. The magnitude of this scandal was revealed to the public in 2018, and its repercussions continue to resonate in the Netherlands to this day. The severity of the situation led to the resignation of the Dutch cabinet in 2021, as it grappled with the ramifications of this ongoing scandal. The scandal encompassed a series of problematic governmental actions, harsh rules and policies, unjustified accusations of fraud, relentless benefit recovery measures, impeding legal and investigative procedures, inadequate and incorrect information, no transparency of the childcare fraud system, and lack of responsiveness by Dutch authorities to the individuals who voiced concerns [4,6,9].

A minor administrative error in applications or renewals, such as missing signatures on childcare service contracts or delayed payment of mandatory personal contributions, were sufficient grounds for being unjustly accused of fraud by the self-learning decision-making algorithm. Consequently, parents and caregivers experienced severe consequences, including substantial repayment obligations, and were labeled as fraudsters. This gave rise to detrimental financial challenges, encompassing debt accumulation and unemployment, thereby impeding their ability to meet rental or mortgage obligations [6,9,12]. Moreover, individuals were plagued with mental health issues and subjected to heightened stress in their personal relationships, often resulting in divorces and broken homes [21], and more than a thousand children [2] were taken into foster care.

The Dutch tax authorities are currently confronted with a newly imposed fine of €3.7 million by the nation's privacy regulator. The fine stems from multiple breaches of the General Data Protection Regulation (GDPR), the European Union's framework for data protection. Specifically, the tax authorities have been found to lack a lawful basis for processing individuals' data and to have retained such information for a duration exceeding the permissible limit. This penalty serves as an acknowledgment of the authority's failure to adhere to the privacy provisions mandated by the GDPR [3].

Chapter 3: About Black Box - Machine Learning Algorithms

This chapter is dedicated to an in-depth exploration of black box algorithms, encompassing their application domains, developmental methodologies, and the pivotal relationship they share with accuracy, precision, and bias. These concepts hold profound significance as they form the foundational framework for comprehending the efficacy of black box algorithms. The motivation for dissecting this subject arises from the pivotal role that black box algorithms played in the context of the Dutch child welfare scandal. Thus, a robust understanding of black box algorithms becomes a prerequisite for a comprehensive grasp of the circumstances and dynamics at play within the scandal.

3.1 Black Box Algorithms and How they are Used

Machine learning algorithms have gained pervasive utilization across a diverse spectrum of domains and contexts in recent times. In essence, machine learning entails the development of algorithms that enable computers to learn from data and improve their performance over time [69]. One distinctive aspect that characterizes machine learning, and often prompts discussions, is the notion of a "black box."

The term "black box" refers to the opacity of the inner workings of certain machine learning algorithms. Unlike traditional rule-based systems, where the logic and decision-making process are explicitly defined and understandable, black box algorithms operate in a manner that is less interpretable to humans. In other words, the transformation of input data into predictions or decisions is not easily explainable, resembling a sealed black box where inputs go in, and outputs come out, without a clear view of the internal mechanisms at play [44].

The operation of black box algorithms represents a departure from the conventional linear decision paths, often found in rule-based systems. Instead, these algorithms employ intricate patterns and relationships within vast datasets, which can be exceedingly complex and nonlinear. Consequently, while black box algorithms excel at generating accurate predictions or classifications, their mechanisms are often challenging to decipher due to the intricate interplay of multiple factors and variables [44].

The selection of the black box algorithm for the Dutch child welfare context was motivated by its suitability for complex models. Instances characterized by a high degree of complexity, exemplified by intricate architectures like deep neural networks or ensemble models, often entail challenges in elucidating the discrete decision-making procedure due to the considerable interplay among a multitude of parameters[44]. In the specific case of the Dutch Tax Customs and Administration, the integration of black box algorithms stemmed from the intricate nature of the child welfare benefit application process, which involves the exhaustive processing and validation of substantial data volumes. By incorporating black box algorithms, the efficiency of the application process was enhanced, leading to decreased waiting times for approval.

3.2 Black Box Algorithm Development Stages

First and foremost, the problem the black box algorithm is intended to solve, must be clearly defined, specifying the type of task it will perform, such as classification, regression, or clustering. The next step is data collection, which requires gathering a diverse and representative dataset comprising both input variables (features) and the corresponding labels (for supervised learning tasks) or outcomes. To ensure data quality, data preprocessing is necessary. This includes cleaning the data, handling missing values, and performing feature engineering to prepare the dataset for training [48,63].

The process of developing a black box algorithm then moves on to algorithm selection. The appropriate black box algorithm must be chosen based on factors like the nature of the problem, data size, and other requirements. Common black box algorithms include neural networks, ensemble methods, and deep learning models. Once the algorithm is selected, it needs to be trained using the prepared dataset. The data is split into training and validation sets, and the algorithm is trained on the input features and target labels. To optimize the algorithm's performance, hyperparameter tuning is carried out. This involves adjusting the hyperparameters using techniques like grid search or random search [48,63].

After training, the model's performance is evaluated on the validation set using appropriate evaluation metrics, such as accuracy, precision, recall, and F1 score, among others. Addressing bias is an essential step in the process. The model must be examined for any biases and, if necessary, mitigated. Fairness-aware evaluation metrics and bias mitigation techniques are employed for this purpose. If interpretability is crucial, techniques like LIME

(Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations) can be used to explain the model's predictions [63].

Once satisfactory performance is achieved on the validation set, the black box algorithm is tested on a separate test set to evaluate its generalization ability. Upon successful testing, the algorithm is deployed in the desired application or system. Continuous monitoring of its performance in real-world scenarios is crucial for ongoing improvement [48].

3.3 Relationship between Bias, Precision, and Accuracy and Black Box Algorithms

Understanding the relationship between bias, precision, and accuracy in black box algorithms is crucial. These three metrics are interconnected and play a fundamental role in assessing the performance of such algorithms.

Firstly, a definition of each metric is required and afterwards, their relation with the algorithms will be provided. Precision is a metric used to evaluate the performance of a predictive model, particularly in binary classification problems. It measures the proportion of true positive predictions (correctly predicted positive cases) among all positive predictions made by the model. In other words, it assesses the accuracy of positive predictions [64].

In addition, in the context of machine learning, bias refers to the presence of unfair or discriminatory behavior in the model's predictions or decisions. Bias can occur when the model systematically favors or discriminates against certain individuals or groups based on protected attributes such as race, gender, or age [65].

Furthermore, Accuracy is another performance metric used to evaluate predictive models, especially in binary classification tasks. It measures the overall correctness of the model's predictions by calculating the proportion of correctly predicted instances (both true positives and true negatives) overall predictions. The accuracy percentage serves as a pivotal metric that indicates the proportion of correct predictions made by the black box algorithm. This measure evaluates the algorithm's performance in accurately classifying instances within the dataset [64].

The relationship of each aspect with the algorithm is:

- Black Box Algorithms and Bias: Black box algorithms can be more susceptible to introducing bias, especially if the training data is biased or contains discriminatory patterns. The opacity of black box models can make it challenging to detect and address biases, leading to potential unfairness in the model's predictions [66].

- Black Box Algorithms and Precision: Black box algorithms, due to their complexity and opacity, can sometimes achieve high precision by identifying specific patterns in the data that lead to accurate positive predictions. However, this high precision may come at the cost of introducing bias in the model. The model may focus only on the majority group, leading to high precision for that group but low precision for minority groups [66].

- Black Box Algorithms and Accuracy: The accuracy of a black box algorithm depends on various factors, including the quality and representativeness of the training data, the complexity of the algorithm, and the nature of the problem being solved. While black box algorithms can achieve high accuracy by capturing intricate patterns in the data, they may also suffer from overfitting or introduce bias that affects overall accuracy [66].

Lastly, the effects of Bias and Precision on Accuracy percentage referring to the results illustrated in Figure 1 [49] will be discussed below:

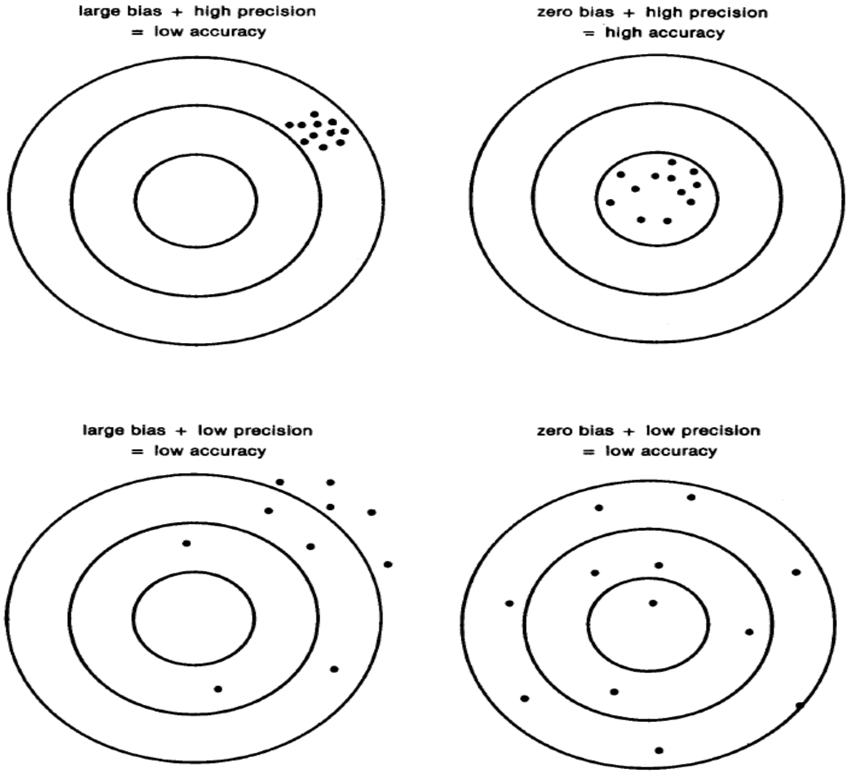

large bias + high precision
= low accuracy
zero bias + high precision
= high accuracy
large bias + low precision
= low accuracy
zero bias + low precision
= low accuracy

Figure 1: Bias, Precision, and Accuracy [Source: Reese, P. (n.d.). *1 Bias, Precision and Accuracy | Download Scientific Diagram*. ResearchGate, Retrieved from: https://www.researchgate.net/figure/Bias-Precision-and-Accuracy_fig2_305767261]

Result in Accuracy Level	Interpretation
Low	Such an algorithm might misclassify instances from certain groups, leading to a low accuracy percentage alongside a pronounced bias outcome.
High	In some cases, a high accuracy percentage does not guarantee that the algorithm is unbiased. The algorithm may achieve high accuracy by correctly classifying instances from the majority group while exhibiting bias against the minority groups. This situation leads to a high accuracy percentage but an unfair bias outcome [50].

3.4 Conclusion: Importance of Cautious use of AI in Decision-making

This chapter has introduced the concept of black box algorithms. It is imperative to recognize that although these algorithms demonstrate remarkable predictive accuracy and efficiency, their limited interpretability can pose a notable drawback, particularly within sensitive domains where comprehending the decision-making process holds paramount importance for ensuring equity, responsibility, and the prevention of biases. It is noteworthy that within the course of examining the available literature, a conspicuous lack of transparency has been observed concerning the Dutch Tax Authority's disclosure of the developmental stages and underlying data associated with such algorithms. Strategies for mitigating biases in black box algorithms will be expounded upon in Chapter 9.

Chapter 4: Conceptually Modeling the Workflow and the Stakeholders

For this research project we used conceptual modeling techniques to provide a better and illustrative representation of the complex child welfare system process.

Models are commonly used for various reasons in different fields and industries. Models help visualize, or picture in human mind, something that is difficult to see or understand [41]. They are simplified representations of real-world systems, processes, or concepts that help scientists understand, analyze, and predict complex phenomena. To represent the complex process of applying for the child welfare benefit, an activity diagram was used.

To represent the complex process of the child welfare system, an activity diagram was designed. This diagram is produced by reverse-engineering and is based on information from literature reviews.

The activity diagram presented in Figure 2 aims at shedding light on the process and facilitating better understanding, despite the limited available information on which it is based.

The model created for this research exhibits certain limitations stemming from a lack of comprehensive explanation of the entire process provided by the government. These limitations will be carefully examined and thoroughly explicated in Section 4.4.

It is worth mentioning that no official model is provided by the Dutch government [42]. Also, it is important to note that the official website of the government's algorithm registry does not contain any information about any algorithm used in the child welfare benefit [42]. Figure 2 shows how models can be useful for better understanding.

4.1 About Activity Diagrams

UML Activity diagrams are instrumental in visually illustrating the flow and sequencing of activities or processes within a system. They offer a clear and intuitive means to comprehend how different activities relate to one another and how they interact. Activity diagrams are particularly valuable for modeling complex processes and workflows, assisting stakeholders in understanding and analyzing the steps required to achieve a specific goal. Such diagrams demonstrate the process from its initiation (the initial state) to its conclusion (the final state), encompassing actions, decision nodes, control flows, start nodes, and end nodes [43].

Understanding the symbols used in the activity diagram is crucial to comprehending the model effectively. Therefore, the table 1 below will present the symbols utilized in the model and their respective meanings.

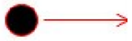
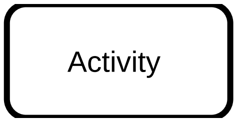
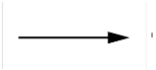
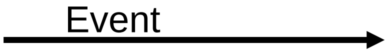
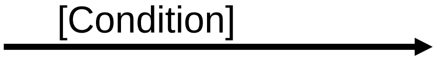
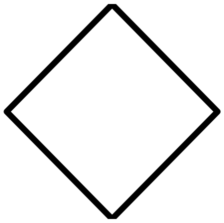
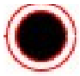
The Symbol	Representation
	**Initial node:** Represents the starting point of an activity.
	**Activity state:** Represents the executable sub-areas of an activity.
	**Control flow:** Represents the flow of control from one action to another.
	**Event :** Represents the event that happens when transitioning from one activity set to the other.
	**Condition:** Represents the condition required to move to the next activity state.
	**Decision node:** Represents a conditional branch point with a single input and multiple outputs.
	**Final node:** Represents the end of all control flows within the activity.

Table 1: UML Activity diagram symbol and their representation (from [43])

4.2 Stakeholders Analysis

Activity diagrams represent the activities of stakeholders in a work process. Table 2 lists the stakeholders involved in the process, along with their respective roles. These are the stakeholders in the so-called Toeslagen Verstrekkingen Systeem (TVS).

Stakeholder Name	Stakeholder Role	Stakeholder Goal
Applicant	Apply for a benefit or for an increase and provide the correct and accurate information in the application. In case of a manual check, provide the correct evidence and information.	Receive the acquired benefit.
Dutch Tax and Customs Administration	Process the application and give the accurate benefit dues to the applicant.	To prevent fraudulent, and large sum of repayment in case of inaccuracy in the application.
AI Agent/Model	Generate a risk score based on the information provided in the application and by the applicant.	To help the tax authorities make neutral decision and prevent fraudulent.
Civil Servant	Perform a manual check for applications with high-risk scores, and contact the applicant to explain the evidence needed. Collect the required evidence and verify it when needed.	To check and verify whether the applicant with a high-risk score is eligible for receiving the benefits or not.
Supervision Team	Handle the objection and complaint application.	Verify the objection and complaint.

Table 2: Stakeholder Roles and Goals in the Child Welfare Benefit.

The workflow of TVS is depicted in Figure 2.

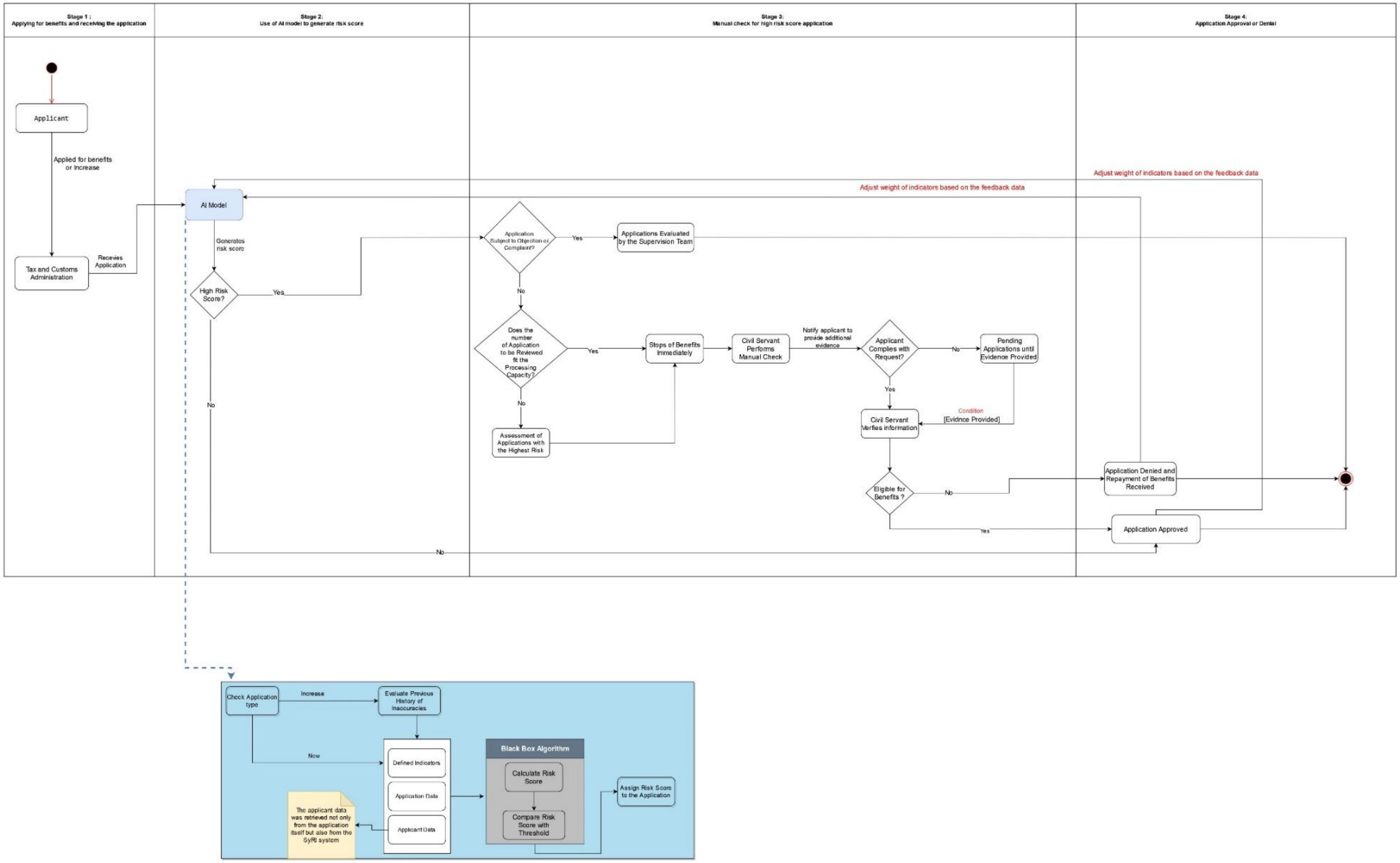


Figure 2: An activity diagram depicts the processes within the workflow of the child welfare application and evaluation by the Dutch tax authorities

4.3 Analyzing the TVS workflow

The Toeslagen Verstrekkingen System (TVS) is an Information and Communication Technology (ICT)-based system specifically designed to manage supplementary benefits in compliance with relevant legislative and regulatory frameworks.

TVS utilizes predefined "rules" to capture citizens' crucial information such as income, household composition, childcare details, and expenditure levels, which are essential in determining their eligibility for benefits.

The risk classification model uses a self-learning algorithm and operates as a black box system to estimate the likelihood of inaccurate benefit applications and renewals [4]. The self-learning algorithm empowers the system to autonomously adapt and evolve without explicit human programming, while the black box nature keeps the internal workings concealed while displaying inputs and outputs [9].

The workflow of the child welfare benefit operation is structured into four distinct stages.

- **In Stage 1,** the applicant submits their request for either a new benefit application or an increase in an existing one, which is received by the Dutch Tax and Customs Administration.
- **Stage 2** involves the utilization of an AI Algorithm to generate risk scores for the applications.
- **In Stage 3**, a manual check is performed by civil servants from the Dutch Tax Administration for applications with high-risk scores.
- Lastly, in **Stage 4**, the applications are either approved or denied.

Each of these stages will be elaborated upon in detail in the following sections.

Stage 1: marks the commencement of the process, initiated when applicants submit their requests for new benefits or seek an increase in existing benefits. These applications are received and processed by the tax and customs administration. Subsequently, the applications are fed into the AI risk classification algorithm, which assesses the provided information for any inaccuracies and incorporates other predefined indicators to generate a risk score.

Stage 2: involves the utilization of the AI model to generate risk scores for the benefit applications. The AI model initiates this process by assessing the application type. In the case

of applications for an increase in benefits, it cross-references them with previous applications to identify any potential inaccuracies. This leads to the assignment of a higher risk score if discrepancies are found in prior applications. Subsequently, both new benefit applications and those seeking an increase are subjected to the black box algorithm, where applicant data and application details, along with predefined indicators, are processed.

The black box algorithm meticulously analyzes applicant data and application information, incorporating various predefined indicators with assigned weights. The process is carried out separately for rent and childcare benefits on a monthly basis. The model considers factors such as income, family situation, and rent/childcare costs, along with additional information like the number of childcare allowance registrations at the applicant's address, for comparison against the model's indicators. Based on this data, the algorithm establishes statistical relationships to evaluate the information against the defined indicators, ultimately generating a risk score and comparing it to certain thresholds. The risk scores fall within a range, with riskier applications receiving scores closer to 1, while less risky applications score closer to 0. In general, the majority of applications yield scores close to 0, leading to an average risk score of 0.05. Additionally, over 90% of allowance applications obtain scores below 0.2, indicating a low level of risk [5].

However, it is important to acknowledge that the applicant data fed to the black box algorithm is not solely derived from the application itself. The model also incorporates data from the SyRI (System Risk Indication) program. SyRI is an ICT system that integrates personal data from various governmental institutions to combat fraud. In early 2020, the district court of The Hague ordered an immediate halt to the SyRI program, with the Ministry later complying with this judgment [51]. A more comprehensive explanation of the SyRI program is provided in Chapter 5 Section 5.2.

Stage 3: After the AI model generates a risk score, low-risk applications are automatically approved by tax authorities [4]. For high-risk applications, an automated selection process is implemented to exclude certain applications from processing and prevent interference with other critical processes, such as complaint and objection handling. These excluded applications then undergo manual evaluation by practitioners from the supervision team [5]. The number of applications to be reviewed manually by civil servants is adjusted based on the available processing capacity, and in case of low capacity, priority is given to those with the highest risk of error.

During the manual check, civil servants carefully examine allowance applications against legal requirements. If inaccuracies are identified, the handler may request additional documents from the applicants [5]. If the applicant complies with the civil servant's request and provides the correct evidence and additional information needed, civil servants verify the information and evidences, along with checking the legal requirements. However, if the applicant does not comply with the request for additional information or evidence, the application will be pending. This procedure can cause delays in payment processing, particularly for new applications, leading to potential financial difficulties for the applicants. For increase benefit applications, the original supplement is paid, while the requested increase remains unprocessed [5].

Stage 4: Once applicants comply with the requests and provide the necessary evidence, civil servants verify the information and evidence, along with checking the legal requirements. If the applicant is eligible, their benefit application, whether for an increase or new application, will be approved. In contrast, if the application is deemed ineligible, it will be denied, and the applicants may face additional consequences, such as repayment of received amounts [4]. The applications that are approved or denied are fed back to the AI model, which serves as additional examples for further model development. The model continuously learns and evolves from these examples, reassessing the relevance and weighting of different indicators. Some indicators may lose predictive value and be removed from the model, while others may gain more influence [4].

4.4 Black Box AI model in TVS

Figure 2 shows how the process involves an AI model, represented in blue, as a "black box" model. This AI model is zoomed in and explained, see the blue box in Figure 2. This model generates a risk score for the application.

In the case of an increase-type benefit application, the system examines previous applications for inaccuracies. Then, both "new applications" and "increase applications", along with their respective applicant details and defined indicators, are inputted into the black box algorithm. This algorithm calculates a risk score, which is then compared against a predefined threshold. Subsequently, an appropriate risk score is assigned to the application. Notably, the applicant data is sourced not only from the application itself but also from the SyRI system. This will be further explored in the following chapter.

4.4 Limitation of the UML Model

The conceptual model in Figure 2 reveals in the child welfare benefit process model a lack of transparency regarding the exact workings of the process. These limitations are primarily manifested in Stage 3, where the manual check is conducted. When the processing capacity is insufficient to review all the applications, priority is given to applications with the highest risk scores. However, there is no clear explanation of how the other applications with lower risk scores are handled in such cases. [5]

Additionally, during the notification phase, when civil servants request additional evidence from the applicants, there is no explicit explanation of the course of action for pending applications in case the applicants fail to comply within a certain period of time. This lack of information raises questions about the disposition of these pending applications and their eventual outcomes. [5]

Chapter 5: Unpacking the System and Identifying Critical Flaws

Upon a thorough examination and analysis of the Dutch child welfare case, it became evident that two systemic challenges were prevalent, giving rise to discriminatory outcomes disproportionately impacting specific segments of the population. These challenges encompass both aspects related to the system's conception and implementation, which will be the focal point of this chapter's discussion, as well as issues pertaining to governmental actions and the transparency of the system, which will be expounded upon in the subsequent chapter.

5.1 The Role of the Black Box AI System in TVS

As discussed in Chapter 3, AI algorithms based on Machine Learning are untransparent to humans, in the sense that the path of decision-making is not explainable, making ML inherently untransparent. [63].

The TVS used an AI system for risk calculation of applicants. To do so, the AI system was "tuned", by feeding it with input. According to [5,8,9], various inputs were fed into the TVS.

Firstly, the AI ML system was trained on data from various sources. To do this database from various sources were linked, for which personal data was accessed. The first method was the data classification model. The second was called the SyRI (System for Risk Indication). These are discussed in the following sections.

Secondly, a risk classification model was used, to identify potential risk based on certain human characteristics. These two inputs to the AI system are discussed below.

5.2 The SyRI System: using Big Data for Human Profiling

The TVS involves the utilization of an AI model that not only relies on the information provided in the application (applicant) itself, but also draws from the SyRI (System Risk Indication) system, a sophisticated big data analysis system. SyRI is granted access to a vast array of data sources, including work records, fines, penalties, taxes, property ownership, housing details, educational records, retirement status, debts, benefits, allowances, subsidies, permits, and exemptions, among others. This data is sourced from various public authorities, such as the Dutch Tax and Customs Administration, the Labour Inspectorate, and the Public Employment Service, creating a comprehensive compilation of citizen data stored by multiple institutions [51].

The SyRI system employs a cross-referencing approach to analyze the collected data. For example, tax information can be compared with data on individuals receiving state aid and support, or information on residential addresses can be aligned with data from the naturalization authority. Through the evaluation of specific risk indicators, the software aims to detect an "increased risk of irregularities," thereby identifying potential instances of non-compliance with the law. For instance, SyRI may raise an alert if an individual receives housing benefits but is not registered at the corresponding address. In response, a designated employee from the Ministry of Social Affairs would conduct a thorough examination of the data. If any concerns arise from this scrutiny, a "risk report" is generated and forwarded to the relevant authorities for further action. Subsequently, the agency overseeing housing benefits would deploy an inspector to investigate the flagged "risk address." If the suspicions are validated, the state aid may be reclaimed accordingly [54].

The operation of the SyRI system involves a cooperative association of different state agencies, where the system's output is relayed to an entity known as the "Inlichtingenbureau" (Intelligence Agency). This service agency, primarily tasked with assisting local authorities in determining citizens' eligibility for state benefits, functions as a data processor on behalf of the Ministry of Social Affairs and Employment. The process begins with merging of data at the Inlichtingenbureau, followed by the encryption of personal data and the creation of a separate dataset containing decryption keys for future use. Subsequently, the data undergoes analysis using a risk model within the SyRI system [54].

Notably, the system employs a selective approach to data decryption. Only data records that trigger flags or indicate potential risks are decrypted and subsequently transmitted to the Ministry of Social Affairs and Employment, where the generation of "risk reports" takes place. It is essential to emphasize that the legislation governing SyRI did not mandate individual notification to data subjects upon the submission of a risk report. Instead, the sole requirement was the prior announcement of the commencement of a SyRI project through publication in the Government Gazette. Only if a data subject explicitly sought access to their data, the Ministry would grant such access [54]. The workflow of the SyRI system is visually presented in Figure 3 [54].

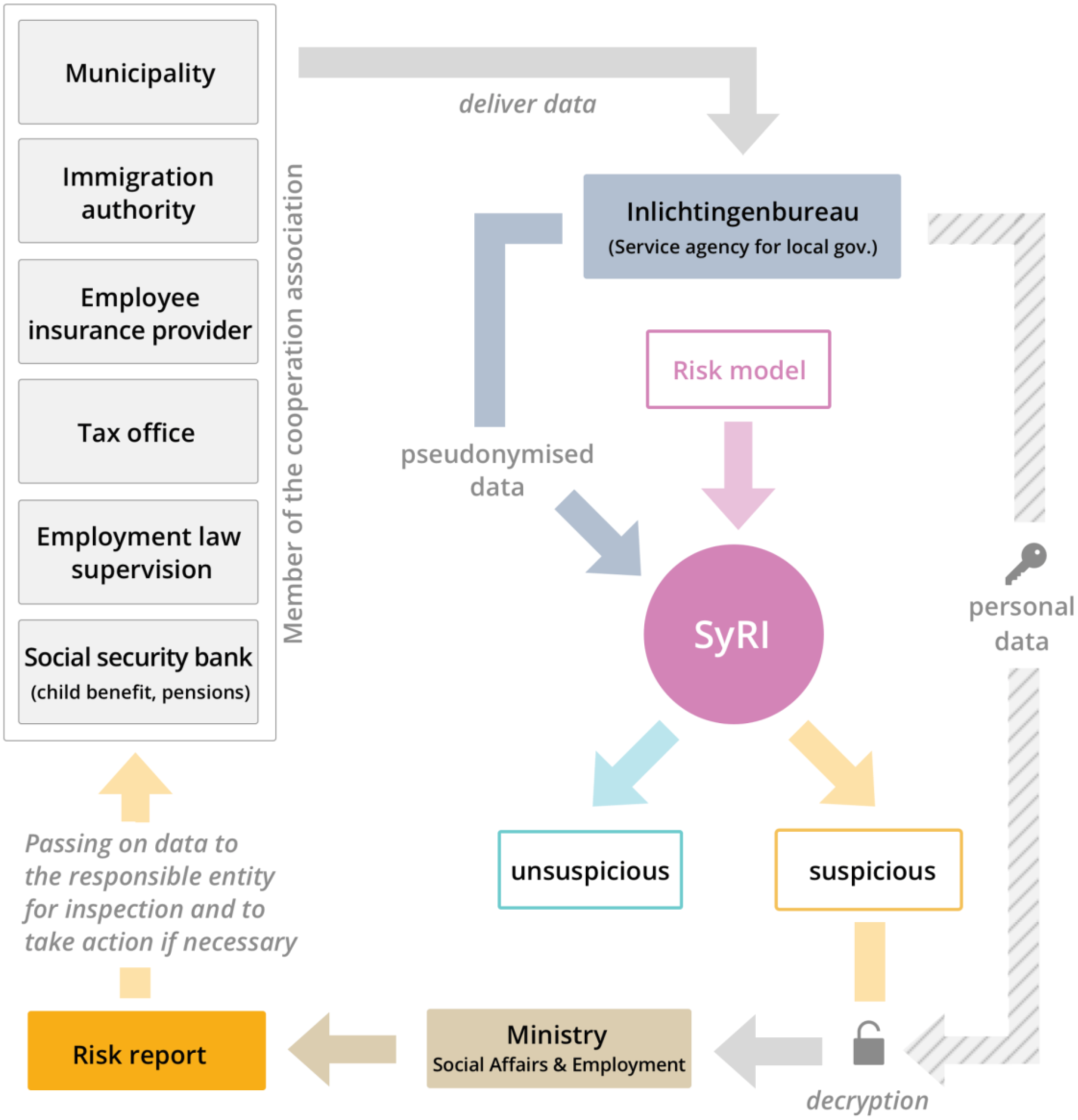

Municipality
Immigration authority
Employee insurance provider
Tax office
Employment law supervision
Social security bank
(child benefit, pensions)
Member of the cooperation association
deliver data
Inlichtingenbureau
(Service agency for local gov.)
Risk model
pseudonymised data
SyRI
personal data
unsuspicious
suspicious
Passing on data to the responsible entity for inspection and to take action if necessary
Risk report
Ministry
Social Affairs & Employment
decryption

Figure 3: How the SyRI system Works in Haarlem [Source: Braun, I. (2018, July 4). *High-Risk Citizens*. AlgorithmWatch. Retrieved from: https://algorithmwatch.org/en/high-risk-citizens/]

5.3 Risk Classification Model Development:

This section delves into the development phase of the risk classification model employed in the child welfare benefit process. Specifically, the focus will be on exploring how the selection of the training sample data and the predefined indicators have implications for bias in decision-making, ultimately leading to discriminatory outcomes, particularly affecting vulnerable citizens.

5.3.1 The indicators used in the Risk Classification Model:

During the period from 2013 to 2019, the childcare allowance model utilized various indicators with distinct weightings to assess the accuracy of applications or changes. These indicators were determined by data analysts and Enforcement employees within Allowances, who considered past experiences and regularly incorporated new examples of correct and incorrect requests into the model. The model aimed to establish correlations, rather than predetermined causality, between indicators and the accuracy of requests. As a result, the values and relationships between indicators could vary each month. Continuous development of the model led to adaptations in the weighting and limit values of indicators, taking into account changing examples and shifts in laws and regulations. Indicators that became irrelevant due to such changes were no longer included in the model's analysis by data analysts [7].

Furthermore, the model considered the interplay and relative importance of indicators rather than assessing them individually. For instance, the number of registrations at an address was linked to the number of children at the same address. This approach accounted for situations where two registrations could represent either a single parent and child or two adults without children. In the context of childcare allowance, the former situation was more common, as the child for whom the parent applied for the allowance was typically registered at the parent's address. The model assessed the probability of errors by examining the correlation between the "number of registrations at address" and "number of children at address" indicators, establishing links between all the offered indicators. Therefore, it is not possible to determine the precise impact of a specific individual indicator on the risk score over time. While the effect of an indicator can be observed, its influence on the risk score itself is not directly discernible. If an indicator is no longer included, the related indicators will respond and display different scores. Consequently, if a new indicator consistently renders an old and less predictive indicator irrelevant, employees would cease offering the redundant indicator to the model [7].

The model analyzed whether the offered indicators were genuinely distinctive in predicting inaccuracies in allowance applications and identified distinctive limit values. For example, the model recognized from benefit applications that a distance exceeding a certain number of kilometers between the home address and the reception address more frequently led to errors in the application compared to a very small distance. Consequently, the model employed a limit value, such as 10 kilometers, for such cases. However, not all indicators held equal levels of distinctiveness. To address this, the model assigned weights to the indicators. The more pronounced the difference between correct and incorrect applications observed for an indicator, the greater its contribution to the risk score. For instance, if a specific value of an indicator frequently appeared in examples of incorrect applications but not in examples of correct applications, the indicator was considered highly distinctive and exerted a significant impact on the risk score. In practice, this was often the case with indicators like the "personal contribution," as errors were more prevalent in applications from citizens making a small personal contribution. It is worth noting that a value could also negatively contribute to the risk score if it had a predictive value for a correct application [5].

Overall, the childcare allowance model utilized a range of indicators with varying weightings and analyzed their distinctiveness to assess the probability of errors in applications. The model dynamically adapted its weighting and limit values based on changing examples and the significance of indicators. By considering the interrelationships between indicators, the model aimed to capture meaningful patterns and enhance its accuracy in evaluating the risk of inaccuracies in allowance applications [5].

Certain indicators utilized in the risk classification model had a significant impact and resulted in discriminatory outcomes against specific groups within society. These indicators encompassed factors such as nationality, age of the citizen's Dutch Social Security Number (BSN), status as a single parent, The Fraud Signaling Facility system (Het systeem Fraude Signalering Voorziening - FVS) , number of children in childcare, distance between the childcare facility and the parent's home, and income level.

5.3.1.1 The indicator Dutch nationality Yes/No:

The consideration of nationality as an indicator in the risk classification model had implications for assessing the risk of inaccurate benefit applications. Specifically, the indicator "Dutch nationality Yes/No" was incorporated into the model to determine whether an applicant was a Dutch citizen. The inclusion of this indicator was motivated by the occurrence of the "Bulgarian fraud"[4], where gangs from Bulgaria were accused of making fake claims for childcare and housing benefits worth 120 million dollars. This fraudulent activity took place during the same period as the model's development, and it coincided with experiences shared by benefits employees who observed difficulties and higher error rates in allowance applications from non-Dutch nationals [5].

The risk classification model assigned higher risk scores to applications from individuals who responded "No" to the "Dutch citizenship Yes/No" indicator, indicating that non-Dutch nationality had a significant impact on increasing the perceived risk in the model. In the context of childcare allowance, it was observed that 21% of the selected group did not hold Dutch nationality, in contrast to only 4% in the total population [8].

The Dutch government's argument that considering citizenship does not equate to using nationality as a risk factor lacks justification. The very fact that individuals were assessed and selected based on their possession or absence of Dutch citizenship implies that nationality influenced the risk-scoring process, at least to some extent. The inclusion of the label "Dutch citizenship: yes/no" played a role in determining the risk score and subsequent treatment by the tax authorities. Although the risk classification model did not differentiate between various nationalities among non-Dutch citizens, it still made a distinction based on nationality [9].

In addition to fraud monitoring, the tax authorities utilized nationality in various other ways within the risk classification model. When suspicions of irregularities or fraud of application had been approved emerged involving individuals with connections to a specific country, civil servants conducted searches in their databases to identify others sharing the same nationalities [10]. This approach aimed to uncover potential cases of organized fraud. Based on previous experiences, the tax authorities employed a range of data, including nationality, family ties, and living conditions [11], to identify larger groups categorized as a single homogeneous population, thereby subjecting the entire population sharing that nationality to suspicion of organized fraud [11]. For example, a fraud alert concerning 120 to 150 individuals with Ghanaian nationality resulted in an investigation of all 6047 applicants with Ghanaian nationality [12]. Regrettably, civil servants made derogatory remarks about families with Caribbean roots, labeling them as an "Antillean nest." [13] The manual selection process utilized by civil servants indicates an underlying assumption linking certain nationalities and ethnic backgrounds to potential fraud, perpetuating negative stereotypes and reinforcing stigmatization.

The utilization of manual selection based on specific nationality in the risk classification model exacerbates the discriminatory effects within the system. This approach follows a "if you look more, you will find more" mentality, as it targets individuals from particular ethnic groups for scrutiny. Consequently, previously unnoticed mistakes or errors are likely to be discovered. These findings are then integrated back into the model, further reinforcing the existing discrimination against non-Dutch citizens.

This approach creates a self-reinforcing cycle of discrimination. By disproportionately scrutinizing individuals from specific ethnic groups, the model is more likely to identify errors or discrepancies in their applications. However, this heightened scrutiny is not applied uniformly across all applicants, leading to an unequal treatment based on ethnicity or nationality. The discovered errors are then used as further evidence to justify the discriminatory treatment, perpetuating the bias within the model.This approach not only reinforces negative stereotypes and stigmatization but also disregards the reality that errors or inaccuracies can arise regardless of a person's ethnicity or nationality.

The combination of targeted manual selection based on ethnicity and the incorporation of the resulting findings back into the model amplifies the discrimination against non-Dutch citizens, contributing to an unfair and biased risk assessment process.

Incorporating the Dutch nationality indicator into the model demonstrates the tax authorities' belief in a correlation between race, ethnicity, and criminal behavior. This acceptance of generalizations based on these factors not only exposes the tax authorities' attitudes towards specific nationalities and ethnic minorities, portraying them in a negative light as deviant or prone to fraud, but also perpetuates the stigmatization of these groups. These practices are in direct contradiction to the prohibition of racial discrimination and underline the necessity of adopting an equitable and unbiased approach.

5.3.1.2 The Age BSN indicator:

The BSN (Burgerservicenummer) is a citizen service number that serves as a personal identification number for individuals in their interactions with Dutch authorities [14]. The age of the BSN was utilized as an indicator in the risk classification model. This indicator indirectly relates to the nationality of an individual, as its description states: "To determine how long someone has been in the Netherlands."[5].

When examining the impact of this indicator on the childcare allowance, a notable distinction arises between the selected group and the entire population. Specifically, the selected group demonstrates an average BSN age of 10 years, whereas the entire population exhibits an average BSN age of 18 years [15].

This finding suggests that even without an explicit Dutch nationality indicator, the design of the model unintentionally revealed protected attributes of the applicants. The inclusion of the BSN age as a factor in the risk assessment process introduces the potential for bias, as it indirectly correlates with an individual's duration of presence in the Netherlands and potentially their nationality. Consequently, concerns arise regarding the potential for discriminatory outcomes based on protected attributes, undermining the principles of fairness and unbiased decision-making.

5.3.1.3 The FVS indicator:

The Fraud Signaling Facility (FSV) system served as a risk signaling mechanism employed by the Tax and Customs Administration to support its supervisory function. Within this system, registrations known as risk signals were generated to indicate potential inaccuracies in tax returns. These risk signals were triggered by various factors, such as instances where individuals claimed significant travel expenses despite residing in close proximity to their workplace. Such circumstances warranted further manual scrutiny of tax returns or allowance requests. Additionally, the FSV system recorded requests for information from other government entities, including municipalities, seeking details regarding an individual's income [16].

FSV played a multifaceted role in the processing of childcare allowance applications, as indicated by internal documents from the Tax and Customs Administration [6]. Furthermore, approximately 10,500 individuals who registered with the Implementing Organization Recovery Allowances (Uitvoeringsorganisatie Herstel Toeslagen - UHT) had their personal data recorded in the FSV. Apart from serving as a blueprint for constructing the risk classification model, FSV also acted as a significant indicator in risk analysis for supplement applications.

Towards the end of 2018, the tax authorities noticed that individuals listed in the FSV were assigned higher risk scores in the risk analysis of supplements. Moreover, a study conducted in November 2019 further highlighted the nourishing effect of FSV entries on the risk estimation process for supplements. This evidence suggests that FSV not only had a role in the training of the model but also continued to influence risk assessments as an independent indicator. As a result, individuals included in the FSV were associated with elevated risk scores in the analysis [6].

5.3.1.4 The Income indicator:

In the risk classification model, income has been acknowledged as a relevant factor by the tax authorities. Historical data used in the model revealed a statistical correlation between income levels and the likelihood of (in)correct benefit applications. The use of income as an indicator occurred multiple times in the Toeslagen model, employing variable limit values since March 2016. Notably, income could have both a positive (higher) or negative contribution to the overall risk score. Lower income was associated with higher risk scores, while higher income resulted in lower risk scores. The selection of income as a factor in the risk assessment process was justified by the Tax authorities as it aimed to prevent high recoveries, which disproportionately affect individuals with low income. Consequently, individuals with lower incomes were more susceptible to being checked for inaccuracies [6].

Despite the acknowledgment of the impact of income on the risk classification, the tax authorities maintain that the influence of the individual variable 'income' on the allowance amount is minimal. They also state that the weight of this variable in the error-checking process is relatively small [6]. However, discrepancies arise when analyzing the data provided by the Tax and Customs Administration. The model assigned a considerably higher risk score to low-income households, as evidenced by figures shared with the Donner Committee. Specifically, of the 1000 highest risk scores generated by the model, 82.3 percent belonged to households with an income of less than 20,000 euros, a figure significantly higher than the 7.3 percent of all benefit applicants falling within the same income bracket [6].

As the exact workings of the model remain undisclosed in public information, understanding the reasons behind this disparity becomes speculative. One possible explanation is the interrelatedness of the indicators chosen by the model. For instance, an indicator based on postal code may indirectly correlate with income, as certain areas are associated with specific income levels. Further research and transparency are necessary to comprehensively comprehend the mechanisms leading to these outcomes and address any potential bias or unintended consequences.

5.3.1.5 Single parent and amount of hours in childcare indicators:

In the risk classification model, two significant indicators that played a crucial role were "single parents" and the "number of people with more than 200 hours of childcare per month." The connection between these indicators becomes apparent upon closer examination. Single parents, who often require more childcare support while they work, tend to have lower household incomes, as the benefits are calculated based on the combined incomes of both partners. Consequently, low income can trigger a high-risk score through various indicators, contributing to potential discrimination in the model's outcomes.

Another factor contributing to the high proportion of low incomes lies in the model's initial data. Officials from the Tax and Customs Administration recently acknowledged that a notable number of low incomes were recorded in the Fraud Signaling Facility (FSV). As the model was trained based on this data, it inadvertently introduced errors into the model, leading it to select on low incomes without a conscious intent to do so [6].

The phenomenon of selection bias comes into play in this context. The model's reliance on data from the FSV, which it perceives as representative of all benefit applicants, creates a distorted picture and leads to a skewed outcome. This becomes problematic when the bias is unconscious, as in this case. Furthermore, the model's usage might have caused a "by selection" effect, resulting in a self-reinforcing selection bias.

5.3.2 Issues of Population and Sample Size for the Risk Model:

Cynthia Liem, an associate professor of Artificial Intelligence at Delft University of Technology, highlights that constructing a risk classification model poses several challenges. In the development phase of the model, around 30,000 examples of both incorrect and correct childcare benefits applications were utilized. These initial examples were obtained from manually assessed applications previously conducted by TVS [6].

The selection of these 30,000 sample files during the development stage significantly influences the model's output. These sample files are essential for training the algorithm, providing it with instances of what constitutes "good" and "wrong" applications. However, it is crucial to acknowledge that the final determination of what is right or wrong is ultimately made by humans, as the algorithm merely examines which indicators align with the labeled good or bad files [6].

Toeslagen, does not fully endorse the quality of these sample files [6]. The files labeled as "good" predominantly pertain to applications that have remained unchanged for an extended period, but this does not necessarily imply the absence of errors or fraud. It is possible that these files were simply not thoroughly examined, highlighting the limitations of relying solely on the labeled sample files for training the risk classification model.

Furthermore, upon conducting a comprehensive analysis, it has been observed that the utilization of a dataset comprising 30,000 examples for training the risk classification model may not be adequate in providing a representative depiction of the entire Dutch population. This becomes particularly crucial when aiming to determine an appropriate sample size that accurately represents a population of 10,083,355 individuals, encompassing adults between the ages of 20 and 64 in the year 2013 [56], and subsequently calculating the risk associated with potential mistakes in their benefit applications. Accomplishing this task necessitates the careful consideration of various factors, including the desired confidence level and the margin of error, to ensure the statistical validity and reliability of the findings.

To assess the adequacy of the existing sample size of 30,000 in effectively representing the broader population of 10,083,355, a meticulous computation of the required sample size must be undertaken based on the specified level of confidence and desired margin of error. This entails utilizing the appropriate sample size formula [57]:

$$n = \frac{Z^{2} * p(1-p)}{E^{2}}$$

where:

- n is the required sample size
- Z is the Z-score, which corresponds to the desired level of confidence (e.g., for a 95% confidence level, Z = 1.96)
- p is the estimated proportion of the population with the characteristic of interest
- E is the desired margin of error or precision

Assuming we want to achieve a 95% confidence level (Z ≈ 1.96) and a 5% margin of error (E = 0.05), and we don't have an estimate of the proportion (p), we can use p = 0.5 to calculate the maximum sample size needed. This provides the most conservative estimate of the sample size.

$$n = \frac{1.96^{2} * 0.5 * (1-0.5)}{0.05^{2}}$$

$$n = \frac{3.8416 * 0.25}{0.0025}$$

$$n = \frac{0.9604}{0.0025}$$

$$n \approx 38416$$

Based on the formula, the maximum sample size required to represent a population of 10.083.355 with a 95% confidence level and a 5% margin of error is approximately 38,416.

Since the calculated sample size is larger than the chosen sample size of 30,000, a sample size of 30,000 may not fully capture the entire population's variability and characteristics. Increasing the sample size to around 38,416 or more would provide a more reliable representation of the population.

5.4 Conclusion of this Chapter: Lack of Control Mechanisms

This chapter has shown several shortcomings of the AI systems used in TVS that could occur through lack of transparency, lack of proper management of the systems and lack of mechanisms for scrutiny. In the next chapter the second type of issue that occurred will be disused which is the governmental issue.

Chapter 6: Governmental Challenges in TVS Model Management and Transparency:

This chapter delves into stakeholder objectives and transparency in child welfare, highlighting instances where stakeholders' alignment with goals was lacking, complicating decisions. The analysis examines the Dutch Tax Customs and Administration's non-disclosure of AI risk scoring, compromising accountability and leaving parents unaware of AI evaluation, potentially undermining fairness [9].

6.1 Unfulfilled Stakeholder Goals

During the child welfare process, several mistakes were observed where stakeholders failed to comply with their goals. Firstly, the Dutch Tax Customs and Administration erred by failing to disclose the utilization of an AI model to generate a risk score for the applicants' benefit applications.

This lack of transparency and disclosure of the AI-driven decision-making process adversely affected the accountability and liability aspects. As a consequence, parents and caregivers were unaware that their cases were being assessed and adjudicated by an AI agent, depriving them of the opportunity to adequately respond or defend their positions. This opacity in decision-making can create a sense of powerlessness and hinder the ability of individuals to participate in the process, potentially compromising the fairness and legitimacy of the child welfare benefit allocation system.

The tax authorities wanted to showcase the effectiveness of the algorithmic decision-making system by successfully recovering ample funds from alleged fraudsters to cover the operation's costs. Consequently, the primary focus was on maximizing the seizure of funds, regardless of the accuracy of the fraud allegations [9].

Secondly, the AI agents exhibited biased decision-making by deviating from their goal of making neutral decisions, instead, displaying bias towards vulnerable groups in society. Further information on this biased decision was provided in chapter 5.

6.2 Transparency and Communication Challenges

The civil servants failed to clearly inform the applicants about the evidence needed to validate their eligibility for benefits. Parents and caregivers were requested to provide additional evidence, but when seeking clarification on specific information considered incorrect or missing, they often encountered a lack of transparency. The civil servants from the tax authorities declined to elucidate their decision-making process, resulting in confusion for the applicants [9].

One of the significant consequences of these mistakes was the false accusation of fraudulent activities, leading to applicants having to repay large amounts of benefits they received. This had a profound impact on society, especially on low-income non-Dutch citizens who were the most affected. Many applicants faced financial difficulties, and some even suffered from health issues as a result of these repercussions. The focus on seizing funds to the maximum extent possible, regardless of the accuracy of fraud allegations, further exacerbated the situation [9].

6.3 Concealment of the Algorithmic Process

The childcare benefits risk classification model was concealed from the public for an extended period, denying parents and caregivers meaningful information about its workings and effects on their individual cases. This lack of transparency hindered their ability to defend themselves and undermined the right to an effective remedy, as well as the principles of good governance and the rule of law.

Using algorithmic decision-making systems in the public sector, without transparency, poses significant risks, including potential discriminatory outcomes [9]. The lack of public knowledge about the existence and impact of such systems creates an information imbalance between affected individuals and those developing and using the algorithms. Transparency mechanisms are crucial to detecting and addressing biases in these systems. Unfortunately, the Dutch tax authorities remained secretive about the childcare benefits risk classification model, hindering oversight by oversight bodies and limiting access to information for journalists, academics, and civil society organizations [24]. The lack of transparency delayed the revelation of the discriminatory practices [9].

Parents and caregivers affected by the scandal were kept unaware that their applications were selected by an algorithm based on high-risk scores for inaccuracy. The tax authorities provided no clear, correct, and complete information about the algorithm's decision-making processes, depriving individuals of the right to meaningful information and obstructing their ability to prove innocence or challenge discriminatory decisions. Moreover, the group most affected by the system's opacity, those with an immigration background, faced additional scrutiny from civil servants, further restricting access to meaningful information. The disclosure of meaningful information would reveal the logic behind the algorithmic decision-making system and its interaction with the civil servant, aiding in the detection of incorrect inputs and outputs, discrimination, automation bias, and the use of black box systems [9].

6.4 Conclusion of this Chapter: Lack of Transparency

This chapter sheds light on the conspicuous deficiency in transparency exhibited by the Dutch government concerning the utilization of black box algorithms in the evaluation of child welfare benefit applications. This absence of transparency has been instrumental in instigating a multitude of challenges that significantly impact vulnerable citizens. The subsequent chapter will undertake an exploration of the extensive ramifications arising from the implementation of such algorithms, thereby unveiling their cascading effects.

Chapter 7: Consequences and Impact

This chapter aims to examine the outcomes of the risk classification model employed in the child welfare benefit system. It will elucidate how the utilization of predefined indicators has resulted in discriminatory outcomes, as evidenced by the cases of participants who have been interviewed for this research and who were subjected to biased decisions by the risk classification model administered by the tax and customs authorities.

Moreover, this chapter will explore the wider repercussions of the algorithm on affected citizens, including the interviewed participants and others who have encountered similar experiences. By analyzing these instances, the chapter seeks to shed light on the impact and implications of the algorithmic decision-making process in the context of the child welfare benefit system.

7.1 Output of the Risk Classification Model

The development of the risk classification model involved a series of significant errors, which subsequently led to discriminatory outcomes against specific minority groups within society. These mistakes encompassed various stages, beginning with the erroneous selection of training data and extending to the biased choices made during the selection of indicators. The manifestation of this discrimination became evident upon a thorough analysis of the model's output.

Between 2014 and 2019, a comprehensive examination was conducted to assess the impact of the risk classification model. The focus of this analysis was on approximately 79,000 unique individuals who underwent manual treatment after being selected by the model. It is important to note that some individuals underwent manual handling multiple times, resulting in a total of around 92,000 treatments. These treatments encompassed approximately 66,000 applications for rent allowance and approximately 26,000 applications for childcare allowance [8].

The State Secretary for Finance – Allowances and Customs conducted an analysis that involved comparing the information from citizens who underwent manual checks by practitioners (referred to as the "selected group") with the entire population of citizens who received a supplement during that period (referred to as the "total population"). To mitigate potential distortions arising from extreme values, measures such as the median, minimum, and maximum scores for specific indicators were taken into consideration. The objective of the analysis was to provide an accurate representation of the "average" benefit recipient [8].

The findings of the analysis revealed a significant discrepancy in the average values of various indicators between the selected group and the total population. This suggests that if a selected citizen had a different characteristic, it is highly likely that they would still belong to the highest risk group due to the influence of other characteristics that contributed to an elevated risk score. Consequently, individuals did not end up in the highest risk category solely based on scoring high on a single indicator [8].

In November 2019, an examination was conducted for the Adviescommissie Uitvoering Toeslagen (AUT) on the 1,000 highest risk scores for childcare benefits using the risk classification model during any month of 2019 [8]. The analysis aimed to provide an overview of these high-risk scores and ascertain the presence of specific characteristics among the selected individuals (as depicted in Figure 2). The findings revealed a significant over-representation of citizens within the group of 1,000 highest scores who possessed one or more identifiable characteristics. This examination further confirms the existence of a distinct group of citizens to whom these specific characteristics apply.

Table 3: highest 1,000 Risk Scores for Childcare Benefits in 2019 [8]

Group	**4 major cities**	**Possess Dutch nationality**	**Family income under €20,000**	**Singel**	**At least 3 children in care**	**Minimum 200 hours of day care**	**Lives more than 10 km from childcare**
1000 highest risk per run	30,8%	78,8%	82,3%	86,9%	12,0%	34,7%	3,8%
Others	12,5%	95,5%	7,3%	14,0%	6,2%	1,2%	0,9%

The analysis of the model's outcomes over the years revealed a consistent pattern of over-representation among citizens who possessed identifiable characteristics. This finding aligns with the analysis conducted for AUT in 2019, as depicted in Figure 1.

7.2 The Chosen Indicators Leading to Discriminatory Results:

Upon closer examination, after understanding how the risk classification model worked and what are the significant indicators that affect the risk score, Participant X, who was interviewed for this study exhibited all the indicators that led to a high-risk score in the risk classification model. Their Dutch Social Security Number (BSN) indicated an age of less than 10 years, signifying the recent acquisition of Dutch nationality. Additionally, they had a low income and three children in childcare, with the childcare facility located more than 10 km from their home. Participant X worked in the health sector, and both they and their partner were employed, with a total of 4 children. During the COVID-19 pandemic in 2019-2020, the participant had to work extra hours, resulting in extended periods of childcare for their child.

For their convenience and better monitoring of their children, they opted to have the childcare facility closer to their workplace rather than their home, leading to the childcare being situated about 10 km from Participant X's home and 2 km from their workplace. However, in 2015, when Participant X first arrived in the Netherlands with their daughter, they were registered as a single parent. In 2017, after the participant's partner joined them in the Netherlands, they received a letter from the tax authority demanding repayment of around 3000 Euros for the single-parent benefit. Later, it was discovered that this was an error in the tax authority's system, as the participant had declared upon arrival that they were not a single parent. Nonetheless, the participant paid the full 3000 euros to the tax authorities.

In 2020, Participant X received another letter from the tax authority, demanding repayment of 8000 Euros, and their childcare supplement was abruptly stopped. Perplexed and seeking clarification, they made multiple contacts with the tax authority but received only instructions to repay the amount promptly. Despite requesting an objection form, the participant did not receive it within the agreed time frame, as the tax employee responsible failed to register the request in the system. The participant was unaware of the 6-week timeframe to apply for an objection, despite various contacts with the tax authorities.

Facing a lack of knowledge about the law in the new country and struggling with the language barrier, Participant X felt compelled to pay the amount in installments to the tax authorities without fully understanding the reasons for repayment. Subsequently, when the participant sought legal assistance, they were informed that they were ineligible for compensation under the childcare scandal. Challenging this decision would entail a lengthy and costly process, making it unfeasible for the participant to pursue further action.

It is worth noting that after the scandal became public, the participant sought the help of a lawyer, paying 150 Euros with the expectation of receiving compensation. However, the government's response was that their situation did not qualify for compensation. The lawyer advised Participant X that challenging this decision would entail significant costs, likely exceeding the potential compensation.

Participant X's case highlights the challenges faced by individuals caught in the complex and potentially discriminatory risk classification model, particularly when encountering language barriers and a lack of legal knowledge in their new country of residence, which unfortunately targeted groups by the risk classification model.

The experiences of Participant X and numerous others who endured the discriminatory outcomes of the risk classification model unveiled the underlying mindset and stereotypes within the Dutch tax authorities. The utilization of nationality in risk scoring reflects the assumptions made by the designers, developers, and/or users of the system that certain nationalities are more prone to committing fraud or engaging in criminal behavior compared to others. Nationality is employed as a defining factor to categorize specific societal groups based on the belief that these groups share common cultural values, traditions, or backgrounds, making them more predisposed to fraudulent or criminal actions. Such differential treatment of individuals by law enforcement authorities in contexts like fraud detection or crime prevention, which lacks objective criteria or reasonable justification and is based on national or ethnic origin, falls under the realm of racial profiling [9].

Racial profiling can manifest covertly, particularly when it is incorporated into algorithmic decision-making systems. In such instances, law enforcement officials need not intentionally differentiate treatment for racial profiling to occur [17]. This phenomenon can arise from either clearly discriminatory attitudes or unconscious bias. Regardless of its form, racial profiling fundamentally violates the principle of non-discrimination. It not only perpetuates historical stereotypical associations between certain categories of individuals and attributes like fraud and ethnic origin but also leads to the criminalization of specific groups, ultimately impacting their well-being and personal dignity [18].

7.3 Consequences of the Dutch Welfare Scandal for Citizens:

The risk classification model had significant and far-reaching consequences and effects on the families that were impacted by its decisions. These consequences included the disruption and fragmentation of families, with many cases resulting in divorces, children being placed in foster care, and in extreme cases, even suicide [19]. Additionally, the model created financial burdens such as bankruptcy, job losses, forced house sales, and homelessness, particularly affecting families from low-income backgrounds. Furthermore, the immense stress and pressure caused by the algorithm's outcomes had adverse effects on the mental and physical well-being of some parents and caregivers, leading to various health issues.

"We were in complete shock and couldn't believe the contents of the letter. When we faced the reality that our only option was to pay the full amount of 8000 Euros to the tax authorities, our lives turned upside down and deteriorated drastically over time," Participant X expressed. According to Participant X, upon receiving the final decision to pay the full amount owed, their partner placed the blame on them for the mistake, as they were responsible for tax matters in the relationship. This led to constant fights and turmoil in their love life. Additionally, due to their low income, both had to work extra hours to repay the owed money, which resulted in heightened stress levels. Eventually, Participant X was diagnosed with a neurological disease caused by extreme stress. "I have spent the last year and a half going from one hospital to another, and from psychologist to another," Participant X shared. This health condition rendered them unable to work, further exacerbating the family's financial problems and stress.

Participant X was not the sole individual impacted by the risk classification model; in fact, it affected around 26,000 families in the Netherlands [20]. These families received claims from the tax authorities, demanding repayment of the allowances they had previously received. In numerous instances, the amount to be repaid amounted to tens of thousands of euros, with some cases exceeding 100,000 euros, leading to severe financial hardship for these families. One such case is that of Franciska Manuputty, a low-income single mother of two, who received a notification in 2010 to repay €30,000 in childcare tax benefits that the government alleged she had not been entitled to. This financial burden left her struggling to pay rent, electricity bills, and even resorting to a food bank to feed her family. Her daughters later disclosed the constant fear they lived in, fearing eviction from their apartment [21].

The algorithms used in the risk classification model were designed to flag "cheats," targeting individuals based on the amount of benefits they received [9]. In other words, the more benefits they received, the more likely they were to fall under suspicion. As a result, those with low-paid jobs, who were most eligible for childcare benefits, were disproportionately affected and came under increased suspicion. Individuals with low incomes heavily relied on income support and benefits to meet their basic needs, and any cuts or suspensions to these benefits led to immediate financial troubles for parents and caregivers from this demographic.

Furthermore, the risk classification model applied a punitive approach, automatically labeling individuals from low-income households as having "deliberate intent or gross negligence" if they received more than €10,000 in benefits or were required to repay more than €3,000 to the tax authorities [12]. Such thresholds disproportionately affected low-income households that would have significantly benefitted from the payment scheme, as they often lacked the financial flexibility to budget for temporary or permanent suspensions of benefits payments, let alone repay large sums of money at once [22].

The detrimental impact of the model was particularly pronounced on people in lower income brackets, a segment of Dutch society that includes a high representation of ethnic minorities. This systemic approach had serious implications for the well-being and financial stability of vulnerable families, exacerbating existing socio-economic disparities.

7.4 Judicial Decision Prohibiting SyRI

In 2020, the SyRI program faced a halt following a judicial decision attributed to its lack of transparency and inadequate data protection measures. The court's ruling was in response to the system's non-compliance with Article 8 of the European Convention on Human Rights (ECHR[1]) and the General Data Protection Regulation (GDPR[2]) [51][52][53]. Additionally, the court expressed apprehension, echoing concerns previously raised by the UN Special Rapporteur, regarding the program's pronounced usage in socioeconomically disadvantaged neighborhoods. Such areas included localities like Capelle aan den IJssel, Eindhoven, Haarlem, and Rotterdam.

[1] Article 8 in ECHR (European Convention on Human Rights) protects your right to respect for your private life, your family life, your home and your correspondence (letters, telephone calls and emails, for example).

[2] Article 8 - Protection of personal data in the GDPR (General Data Protection Right) states:1.Everyone has the right to the protection of personal data concerning him or her. 2.Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified. 3.Compliance with these rules shall be subject to control by an independent authority.

7.5 Conclusion of this Chapter: Discriminatory Outcomes Affecting Many People

In conclusion, this chapter has delved into the intricate outcomes and consequences arising from the utilization of the risk classification model within the child welfare benefit system. Through an analysis of discriminatory outcomes and their impact on citizens, a deeper understanding of the challenges and complexities of automated decision-making has emerged.

Upon further analysis, an intriguing relationship between the civil servants responsible for conducting manual checks on applicants and the utilization of automated decisions has become apparent. This interplay has significant implications for the decision-making process of civil servants. As the risk classification model generates scores that subsequently influence the manual handling of applications, civil servants may unwittingly be influenced by the automated output. This can lead to a reinforcement of biases, as automated decisions could subconsciously shape the perspectives of civil servants, even when they are tasked with manual evaluations. This topic will be explored in greater detail in the following chapter.

Chapter 8: Human Empathy and Automated Decision Making:

In the realm of government operations, a broad spectrum of administrative functions, ranging from granting licenses to disbursing payments and overseeing claims, is carried out by governmental entities. Traditionally, these tasks have been undertaken by human personnel. However, the evolving digital landscape is gradually leading to an increased reliance on automation for government operations. The inherent potential of automation lies in its capacity to enhance efficiency, speed, and accuracy. Nevertheless, even in the context of an automated state characterized by responsibility, impartiality, and integrity, a crucial element of effective governance appears to be absent: the human element [58].

Efficiency and precision are indeed inherent advantages of automation. Nonetheless, a pivotal aspect of good governance, the empathetic connection with citizens, seems to be overshadowed in this automated scenario. The initial perception of a digitally-driven governmental future appears somewhat sterile, lacking the crucial quality of empathy. Empathic interaction with the public remains an indispensable facet of any responsive and accountable government authority. In an increasingly digitized era, ensuring that individuals have avenues to engage with human representatives, express their perspectives, and have their concerns acknowledged remains a fundamental duty of government [58].

Against this backdrop, the rise of semi-automated and automated decision-making systems prompts a critical examination of their impact on human empathy and the legal accountability of civil servants involved in decision-making processes. It is imperative to investigate how the utilization of such systems could potentially influence civil servants' sense of duty and compassion, thereby shaping their attitudes and conduct, particularly within the context of child welfare benefit administration.

8.1 Diminished Sense of Responsibility:

The incorporation of semi-automated and automated decision-making systems may lead civil servants to perceive themselves as mere executors of algorithmic outputs rather than active decision-makers. When decisions are largely dictated by predefined indicators and algorithms, civil servants may feel less accountable for the outcomes, as the final judgment rests with the system rather than their individual discretion. This diminished sense of responsibility can lead to a detachment from the decision-making process and reduce the motivation to critically evaluate the decisions made by the system [60].

8.2 Impact on Empathy and Human Connection:

Civil servants traditionally play a crucial role in interacting with welfare benefit applicants, understanding their unique circumstances, and applying empathy and discretion when assessing applications. The introduction of semi-automated and automated decision-making systems may shift the focus away from personal interactions, potentially reducing the opportunity for civil servants to empathize with the applicants. As a result, the human connection and compassionate understanding that civil servants once provided may become diluted in favor of strict adherence to the system's decisions [60].

This shift is evident in instances such as the child welfare scandal, where minor administrative oversights in applications or renewals, such as the absence of signatures on childcare service contracts, served as sufficient grounds for unfounded fraud allegations. In this scenario, applicants found themselves subject to stringent regulations, rigid interpretations of laws, and ruthless benefits recovery policies despite trivial administrative errors.

8.3 Perceived Technological Determinism:

The use of automated decision-making systems may contribute to a perception of technological determinism among civil servants, wherein they believe that the algorithm's decisions are inevitable and beyond their influence. This perception may lead to a disempowerment of civil servants and a sense of fatalism regarding the outcomes, as they may believe their individual efforts to advocate for applicants could be futile against the authority of the algorithm [59].

8.4 Liability Concerns:

To mitigate the potential negative effects on civil servants' sense of responsibility and empathy, the following measures can be considered:

1. **Training and Education:** Providing comprehensive training to civil servants about the functioning and limitations of the automated decision-making system can help them understand their role in the decision-making process and maintain a sense of responsibility for the outcomes [60].

2. **Human Oversight:** Incorporating human oversight in the decision-making process can create a balance between automation and human discretion, allowing civil servants to retain a degree of agency and empathy in their interactions with applicants [60].

3. **Ethical Guidelines:** Establishing clear ethical guidelines for civil servants to adhere to when implementing decisions made by the system can reinforce their accountability and commitment to empathy in their work [60].

The implementation of semi-automated and automated decision-making systems within the child welfare benefit framework raises considerations regarding the impact on civil servants' empathy and accountability. Recognizing that empathy is a cornerstone of effective governance, it is essential to understand that citizens who are subject to decisions that could significantly impact their lives should have the avenue to interact with human representatives. The ability to communicate their situations and concerns contributes to a more responsive and compassionate governmental structure.

8.5 Conclusion of this Chapter: Fostering Empathy and Responsibility in the Welfare Benefit System

By carefully addressing the potential challenges posed by automation and advocating for a harmonious blend of automated processes and human involvement, it is feasible to uphold a welfare benefit system that remains both empathetic and responsible. This approach takes into account the indispensable role of civil servants in upholding fairness and empathy within the decision-making process, thereby ensuring that the needs of applicants are adequately met while preserving the principles of accountability and compassion.

Chapter 9: Discussion and Recommendations

This chapter will provide recommendations to mitigate the issues previously discussed in chapter 5 and 6 to prevent their recurrence in future implementations.

9.1 Risks of Black Box Systems and Self-Learning Algorithms in Public Sector:

Black box systems are algorithmic systems that lack visibility to users and other parties, hindering oversight, accountability, and transparency [9]. The risk classification model used a black box system, where applications with high inaccuracy scores underwent manual checks by a civil servant, who lacked access to information about the basis for risk scores [4]. This hindered meaningful accountability and transparency in the tax authorities' fraud detection practices.

The use of self-learning algorithms in the public sector poses significant risks to human rights, good governance, and the rule of law. These algorithms continuously adapt decision-making based on new information, reaching a point where no human, including designers and developers, can fully understand their decisions [23]. This lack of verifiability and predictability undermines government activities, and there is a risk of amplifying biases due to erroneous assumptions in the self-learning process [9]. Such characteristics are incompatible with good governance and the rule of law, especially when used in decision-making impacting individuals' rights and society [9].

9.2 Modeling Techniques for the Sake of Openness:

The activity diagram model, which was introduced earlier and is derived from publicly available data, stands as a compelling illustration of the transformative potential of modeling in enhancing transparency. This significance becomes particularly apparent when considering instances such as the recent UWV scandal. In this case, the state benefits agency UWV was discovered to have illegally employed a system for gathering information about unemployment benefits and assessing whether claimants might be residing abroad [67]. Furthermore, the enigmatic nature of the algoritme.overheid.nl [42] website further underscores the Dutch government's proclivity for non-disclosure. Despite this trend, the activity diagram provides a tangible counterpoint, showcasing that even rudimentary modeling techniques can offer remarkable clarity. Such visual representations can prove invaluable in elucidating complex processes for a diverse range of stakeholders, including civil servants, applicants, and tax authorities.

For civil servants tasked with navigating intricate decision-making systems, an activity diagram offers a comprehensible visual aid. It provides a step-by-step breakdown of the decision process, elucidating how automated algorithms and human intervention interact to reach conclusions. This enhanced understanding can empower civil servants to better comprehend their role within the system, fostering a heightened sense of responsibility and accountability.

Likewise, for applicants seeking welfare benefits, the activity diagram offers a transparent overview of the evaluation process. It demystifies the intricate stages involved in assessing applications and highlights potential pitfalls that might lead to adverse outcomes. By demarcating the journey from application submission to decision, modeling provides applicants with a clear map of the process, enabling them to engage more effectively with authorities and challenge decisions when necessary.

Furthermore, the benefits of modeling extend to the tax authorities themselves. By visualizing the entire process, tax authorities gain a holistic view of the decision-making mechanism. This vantage point can reveal potential inefficiencies, bias points, or areas where human intervention may be lacking. Armed with this comprehensive understanding, authorities can refine their processes, ensuring a fair and transparent system.

In light of these advantages, it is undoubtedly prudent for the government to consider modeling as a recommendation to enhance transparency. Visualization not only bridges the gap between intricate algorithms and stakeholders but also facilitates communication, accountability, and informed decision-making. The activity diagram stands as a testament to the potential of modeling in fostering transparency and promoting effective governance within complex decision-making frameworks.

9.3 Safeguarding Algorithmic Decision-Making

The Dutch tax authorities did not conduct a human rights impact assessment before implementing the childcare benefits risk classification model, leading to the use of discriminatory algorithms without proper mitigation measures. To address such risks, it is essential for states to continuously assess and monitor the human rights impact of algorithmic decision-making systems throughout their lifecycle and take appropriate mitigation measures. The lack of a human rights impact assessment in the childcare benefits scandal resulted in harsh treatment for certain groups of people, violating their human rights. Implementing a mandatory and binding human rights impact assessment for public sector use of algorithmic decision-making systems, involving relevant stakeholders and independent human rights experts, is crucial to ensure the respect, protection, and fulfillment of human rights.

9.4 Strengthening Human Rights Oversight:

The oversight on the tax authorities' use of the childcare benefits risk classification model was ineffective due to the fragmentation of existing oversight mechanisms, lack of binding human rights oversight, and the tax authorities' opaque operations. Various institutions, such as the National Ombudsman in 2017 [25] and the Netherlands Institute for Human Rights in 2020 [26], examined the practices, but their oversight was not comprehensive or legally binding. The Dutch Data Protection Authority's investigations were hindered by misleading information from the tax authorities, which led to a flawed approach in assessing discrimination [27]. The Dutch Data Protection Authority was unable to determine whether the tax authorities had processed data points on ethnicity, such as race or skin colour, and argued that (in general) nationality should not be considered a direct proxy for ethnicity or race data [28]. This underscores the need for a dedicated human rights oversight body with binding powers to scrutinize algorithmic decision-making systems and their impact on human rights. Governments should establish independent and comprehensive human rights oversight mechanisms to strengthen accountability and protect human rights in the use of algorithmic decision-making systems in the public sector.

9.5 Ethnicity and Nationality in Law Enforcement Risk Profiling:

The Dutch government allows the use of ethnicity and nationality as risk factors in law enforcement decision-making, despite publicly opposing racial profiling. This practice has been met with debate and criticism, with Prime Minister Mark Rutte refusing to discontinue the use of nationality in risk-profiling [29]. The government's stance on this matter has not changed despite calls to prohibit such discriminatory measures. The use of nationality and ethnicity in risk-profiling leads to racial profiling and disproportionately affects individuals from ethnic minority backgrounds. The Dutch Human Rights Institute has highlighted this issue in its publication on racial profiling[9]. To address this concern, the Dutch government should enact a clear and legally binding ban on the use of nationality and ethnicity data in risk scoring for law enforcement purposes when there is no individualized suspicion of wrongdoing.

9.6 Seeking Justice in Algorithmic Systems:

The lack of transparency and accountability within the tax authorities led to a complex legal process for the victims of the childcare benefits fraud system. Remedial actions taken by the Dutch government failed to address discrimination caused by the risk classification model. International law requires that individuals have the right to an effective remedy and adequate redress for human rights violations [30], but victims often face challenges when seeking justice in the context of algorithmic decision-making systems. The opacity and constant adaptation of these systems, to the extent that even their designers may struggle to explain outcomes [31], hinder individuals' understanding of their impact and access to remedies. Good governance principles play a crucial role in realizing human rights in the data-driven society, requiring states to investigate algorithmic biases and impose sanctions when necessary.

The childcare benefits scandal presented numerous obstacles for victims to access justice and effective remedies. The tax authorities' lack of transparency and refusal to provide information further hindered redress efforts. Parents and caregivers identified as fraudsters received no explanations for years [32], and requests to inspect files were often denied [33]. Complaints filed with the tax authorities took two years to be handled [34]. While some remedial measures were eventually taken, they excluded discrimination based on nationality, ethnicity, or social origin resulting from the risk classification model. Currently, there are no effective remedies for racial profiling and other discrimination caused by the algorithmic system [35].

To address these issues, states must ensure meaningful accountability and provide effective remedies for human rights harms related to algorithmic decision-making systems. This includes creating independent and accessible processes for redress and designating roles within the public sector responsible for timely remedies, subject to accessible appeal and judicial review.

9.7 Mitigate Bias in Black Box Algorithms:

Eliminating bias in black box algorithms presents a challenging task due to their inherent lack of transparency, which makes it challenging to directly discern their internal decision-making processes.

Ensuring the training data used for algorithm development is diverse and representative is pivotal. Through meticulous curation and addressing imbalances, biases inherent in the data can be minimized, leading to increased fairness in algorithm predictions. Although black box algorithms might not offer direct insights into their decision-making, various techniques can be employed to detect and assess bias in their predictions. Employing fairness-aware evaluation metrics and proxy models that approximate decision-making processes can aid in identifying bias and locating potential avenues for improvement.

Researchers have devised a range of bias mitigation techniques applicable post-training to reduce bias in algorithmic predictions. These methods could involve data reweighting, loss function modification, or the application of fairness constraints during optimization. Furthermore, integrating interpretability techniques alongside black box algorithms can provide valuable insights into predictive influences. Techniques such as feature importance analysis, partial dependence plots, and LIME (Local Interpretable Model-agnostic Explanations) can pinpoint potential bias sources and enhance comprehension of algorithm behavior. Ensemble methods present another promising solution by amalgamating multiple models, including interpretable ones, to balance predictive accuracy and interpretability. Ensuring interpretable models exert a stronger influence on crucial decisions can mitigate potential biases from black box models. Additionally, incorporating fairness-aware regularization terms during training can guide algorithms to learn equitable representations and reduce bias, further enhancing fairness [48].

Continuous monitoring of black box algorithm performance is imperative post-deployment. Regularly assessing predictions and outcomes enables swift detection and rectification of emerging biases, ensuring system fairness over time. It is vital to acknowledge that, given their inherent lack of transparency, complete elimination of bias in black box algorithms may remain unattainable. Nonetheless, endeavors should be directed towards striking a harmony between predictive accuracy and fairness. Whenever feasible, the pursuit of transparency in decision-making should be prioritized to ensure ethical and responsible utilization of AI systems [50].

Lastly, ethical considerations hold a pivotal role in black box algorithm development and deployment. Upholding transparency, fairness, and accountability must remain paramount to ensure responsible usage and circumvent unintended consequences. By embracing a dual focus on accuracy and bias, developers and researchers can contribute to the creation of more responsible and equitable black box algorithms, ultimately benefiting society while minimizing potential biases [50].

9.8 Need for Transparency

The use and development of black box algorithms were discussed because throughout the study of literature reviews, there has been a lack of disclosure regarding the stages or steps taken to develop the black box algorithm for the child welfare benefit. The absence of this information raises concerns about transparency by the government in the development and implementation of such algorithms.

Transparency is of paramount importance in the context of government practices, especially when dealing with sensitive matters like social welfare benefits. When the government is transparent about the process of developing and implementing black box algorithms, it fosters accountability and trust among the public. Citizens have the right to understand how such algorithms work and how they may impact their lives.

Transparency also allows for external scrutiny and evaluation of the algorithm's fairness and potential biases. When the government is open about the methods used and the data sources involved, it enables independent experts, researchers, and advocacy groups to assess the algorithm's performance and identify any unintended biases or discriminatory outcomes.

Moreover, transparency ensures that citizens have access to essential information about the decision-making process, which promotes fairness and equity in the distribution of benefits. When individuals know the criteria used to assess their eligibility for welfare benefits, they can better understand the outcomes and have the opportunity to challenge decisions if they believe errors or biases have occurred.

Disclosing the stages and steps involved in developing black box algorithms for child welfare benefits is crucial for ensuring transparency by the government. Transparency promotes accountability, trust, and fairness in algorithmic decision-making, benefiting both the government and the citizens it serves.

To ensure transparency, governments should create public registries with comprehensive information on the use of algorithmic decision-making systems in the public sector. They should provide affected individuals with meaningful information about the logic and consequences of decisions, even if human intervention is involved in the process. This transparency is vital to protect human rights, promote good governance, and address potential biases in algorithmic systems. By employing black box algorithms, the application process becomes more efficient, reducing waiting times for approval.

The use of black box algorithms has the potential to mitigate human biases, as the decision-making process can be more neutral compared to human judgment. However, it is crucial to ensure that the algorithm is developed and implemented correctly to prevent transferring human biases to the machine. While machines themselves are not inherently biased, any biases present in the data or design can influence their decisions, thus making it biased. Therefore, it's very crucial when implementing a black box algorithm in public administration to have continuous supervision of their decisions to ensure that no biases are introduced later.

9.9 Achieving Adequate Representation in Training Data

As discussed in Chapter 5, the training data used in the risk classification model for child welfare was insufficient and failed to represent Dutch society as a whole. Therefore, ensuring the representativeness and reliability of the sample files becomes paramount for the development of an effective AI model. The selected samples must encompass a diverse array of scenarios and accurately mirror the intricacies and nuances found in real-world situations. If the sample files fail to encompass the full spectrum of potential errors, fraud, or inaccuracies, the risk classification model might produce erroneous outcomes, unable to accurately identify high-risk applications. Hence, a meticulous approach to selecting and validating sample files is essential to bolster the model's robustness and reliability.

Moreover, the engagement of human experts in the development and training process holds significant importance. Their specialized knowledge empowers them to evaluate the precision and soundness of the sample files, ensuring alignment between the risk classification model and its intended objectives and regulatory frameworks. Human oversight and validation assume a pivotal role in the AI system's overall performance and accountability, serving as a crucial safeguard against potential biases, errors, or unforeseen ramifications.

In summation, the meticulous choice of sample files and the utilization of a broader and more diverse training dataset stand as pivotal measures during the creation of a risk classification model. These determinations exert substantial influence over the model's capacity to precisely classify and evaluate applications for supplementary benefits, thereby upholding fairness and transparency throughout the process. By striving for inclusive and impartial data, developers can forge a model that honors the varied needs and circumstances of the entire Dutch populace.

9.10 Need to Comply to Data Protection Rules (GDPR)

During the examination of the child welfare benefit system, a discernible pattern emerged where applications classified as low risk received automatic approval without undergoing human scrutiny [4]. In the context of these cases, the risk classification model operated as a fully automated decision-making mechanism. It is notable that such fully automated decisions are deemed prohibited according to Article 22 of the General Data Protection Regulation (GDPR) [61]. Whether the effects of these automated decisions were positive or negative, individuals affected by them possess the entitlement to substantial insights into the algorithmic rationale underpinning the risk classification model.

Additionally, in situations where human intervention is present, such as in the case of applications rejected by the risk classification model, the General Data Protection Regulation mandates that individuals affected by semi-automated decisions are entitled to access meaningful information. This meaningful information would unveil the logic underpinning the algorithmic decision-making system and the interaction between the system and the civil servant [61,62].

However, it is unfortunate that the right to meaningful information is not extended to parents and caregivers who have been subjected to decisions that carry discriminatory implications. In such instances, these individuals are denied access to the comprehensive information that would empower them to grasp the reasoning and intricacies of the decision-making process.

Another challenge lies in the way AI processes model individuals based on quantifiable characteristics, reducing them to predefined categories. AI mathematical processes model the world as black and white. The process of representing an individual as a cluster of quantifiable characteristics using a vector pushes individuals into categories [39]. This approach fails to capture the uniqueness and individuality of people, as AI decision-making evaluates individuals against an optimized combination of characteristics, overlooking their diverse experiences and circumstances. As a result, some individuals may be denied resources or opportunities simply because they don't fit the algorithmically constructed profile of a "good applicant," or their unique aptitudes are not recognized by the algorithmic processes.

In the context of the childcare benefits scandal, the group that arguably stands to benefit most from transparency—parents and caregivers—found themselves largely excluded from accessing meaningful information pertaining to the algorithmic logic. The inability to access information about the logic behind the algorithm raises concerns about the transparency and fairness of the decision-making process, particularly with regards to the interaction between the automated system and human intervention. This interaction plays a pivotal role in identifying and rectifying incorrect inputs and outputs, addressing potential instances of discrimination and automation bias, and uncovering the inner workings of opaque "black box" systems [39].

9.11 Need for Procedures of Ethical Clearance

Within the context of AI in public administration it is akin to constructing a robust protective barrier. This ethical barrier helps prevent unfairness and discrimination. Ensuring that the AI models used are honest and unbiased from the outset is crucial. This involves carefully examining different aspects of the AI, such as its functionality and information sources. By conducting this careful scrutiny, it can be ensured that the AI is built on a strong ethical foundation, benefiting individuals and avoiding harm [68].

When creating new AI systems for government tasks, it's necessary to ensure they adhere to ethical rules. This means conducting thorough checks to determine if the AI treats everyone fairly and acts responsibly. This way, it can be ensured that AI in government is not only about technology but also about fairness and doing what's best for people. This fusion of ethical consideration and technological progress lays the groundwork for AI to positively impact society while upholding moral responsibility.

When artificial intelligence (AI) is introduced in fields such as government activities, considering ethics becomes crucial. For example, in the case of child welfare benefits, it's evident that those responsible for developing the system didn't give sufficient thought to what's right and wrong when using these AI models. This highlights the insufficient attention paid to ethical considerations in AI usage, underscoring the need to thoughtfully ponder ethical principles and fair treatment when employing AI in government tasks. To ensure fair usage of AI and its benefits for all involved parties, it's essential to establish clear ethical rules right from the beginning of AI integration.

Reflecting on ethics within the context of AI is akin to constructing a robust protective barrier. This ethical barrier helps prevent unfairness and discrimination. Ensuring that the AI models used are honest and unbiased from the outset is crucial. This involves carefully examining different aspects of the AI, such as its functionality and information sources. By conducting this careful scrutiny, it can be ensured that the AI is built on a strong ethical foundation, benefiting individuals and avoiding harm [68].

Furthermore, especially when creating new AI systems for government tasks, it's necessary to ensure they adhere to ethical rules. This means conducting thorough checks to determine if the AI treats everyone fairly and acts responsibly. This way, it can be ensured that AI in government is not only about technology but also about fairness and doing what's best for people. This fusion of ethical consideration and technological progress lays the groundwork for AI to positively impact society while upholding moral responsibility.

9.12 Conclusion of this Chapter: Concluding Recommendations

In conclusion, the discussion and recommendations presented in this chapter shed light on the multifaceted challenges of implementing algorithmic decision-making systems in the public sector. The exploration of risks posed by black box algorithms, self-learning systems, and discriminatory practices underscores the importance of transparency, accountability, and ethical considerations.

The proposed recommendations emphasize the need for comprehensive oversight, transparent modeling techniques, representative training data, adherence to data protection rules, and ethical clearance procedures. These measures collectively aim to safeguard human rights, mitigate biases, and promote fairness within algorithmic systems.

By embracing these recommendations, governments can pave the way for responsible and ethical integration of AI into public administration. Striking a balance between technological advancement and fundamental values of transparency, equity, and accountability is crucial for creating a future where AI-driven decision-making serves as a tool for positive societal transformation.

Chapter 10: Conclusion

In light of the extensive insights gathered throughout this research, it is evident that the use of AI algorithms by public authorities, while holding immense potential, also presents formidable challenges that demand immediate attention. The rapid integration of AI within government agencies has raised apprehensions, chiefly due to the absence of thorough risk assessments prior to deployment. Central to these concerns is the opaqueness inherent in machine learning models during the decision-making process. This opacity obscures the rationale behind decisions, obstructing the essential scrutiny necessary for ensuring fairness, impartiality, and transparency, while simultaneously blurring lines of accountability. In a democratic society, this predicament poses a significant risk, undermining the very foundations of governance.

The overarching objective of this research has been to dissect the current landscape of algorithmic decision-making systems within public administration and unravel the potential risks when such systems are entrusted with shaping human lives. The inquiry has been anchored in a thorough examination of the Dutch child welfare scandal—a poignant case that unfolded between 2004 and 2019. During this period, the Dutch government's tax and customs administration deployed AI algorithms to identify potential cases of fraud among citizens, leading to the wrongful accusation of around 26,000 individuals. The aftermath of these inaccurate accusations cast countless citizens and their families into a distressing quagmire of social, emotional, and financial hardships that endured for years.

Employing a methodological framework that leverages conceptual modeling, this research scrutinized workflows and decision-making processes, effectively peeling back the layers of complexity to expose latent biases and unethical decision pathways. This analytical approach not only unveiled the intricate machinery at play but also illuminated areas of concern that were previously shrouded in obscurity.

The value of this research transcends its examination of the Dutch child welfare scandal; it provides a seminal contribution to comprehending the potential hazards intrinsic to the utilization of AI in public administration. Furthermore, it introduces a potent tool in the form of conceptual modeling techniques, harnessed from the realm of computer science, to untangle and elucidate intricate work processes. Beyond mere analysis, this study proffers a series of recommendations that hold the promise of imbuing AI-supported decision-making within public administration with newfound fairness and transparency.

The significance of these findings resonates deeply, especially considering the cautionary narrative of the Dutch child protection services. It serves as a clarion call for the imperative need to imbue AI systems with qualities of explainability, equity, and accountability. Addressing these imperatives necessitates an unwavering alliance between policymakers, computer scientists, and domain experts. Only through collaborative effort can algorithms be meticulously designed to fortify against biases and discrimination, and to erect bastions of transparency, accountability, and human rights within AI-guided decision-making processes. This research, while a single step in a complex journey, sets the trajectory toward a future where AI serves as a force for positive transformation while safeguarding the core tenets of ethical governance.

References:

1. *Applying for child benefit | Child benefit*. (2021, August 10). Government.nl. Retrieved July 5, 2023, from https://www.government.nl/topics/child-benefit/applying-for-child-benefit.

2. “DPG Media Privacy Gate,” n.d. https://www.volkskrant.nl/nieuws-achtergrond/ruim-1-100-kinderen-van-gedupeerden-toeslagenaffaire-werden-uit-huis-geplaatst~baefb6ff/?referrer=https://www.google.com/.

3. Heikkilä, M. (2022, March 29). *Dutch scandal serves as a warning for Europe over risks of using algorithms*. POLITICO. Retrieved July 5, 2023, from https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/.

4. Dutch Data Protection Authority Title: *Report on Tax Authorities' Childcare Allowance (Kinderopvangtoeslag) Investigation* from: https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf , Page. 14 , Date: July 7, 2023.

5. Parliamentary Papers II, 2021-2022, 31 066, no. 923, letter 26 Nov 2021.

6. *Hoe de Belastingdienst lage inkomens profileerde in de jacht op fraude*. (2021, November 22). Trouw. Retrieved July 9, 2023, from https://www.trouw.nl/politiek/hoe-de-belastingdienst-lage-inkomens-profileerde-in-de-jacht-op-fraude~bbb66add/.

7. Parliamentary Papers II, 2021-2022, 31 066, no. 938, letter 8 Dec 2021

8. Letter to parliament on the analysis of the risk classification model Toeslagen, 21 april 2022, Appendix 1.

9. *Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal*. (2021, October 25). Amnesty International. Retrieved June 1, 2023, from https://www.amnesty.org/en/documents/eur35/4686/2021/en/.

10. Dutch Data Protection Authority Title: *Report on Tax Authorities' Childcare Allowance (Kinderopvangtoeslag) Investigation* from: https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf , Page. 19 , Date: July 7, 2023.

11. Dutch Data Protection Authority Title: *Report on Tax Authorities' Childcare Allowance (Kinderopvangtoeslag) Investigation from*: https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf , Page. 21 , Date: July 7, 2023.

12. Dutch Data Protection Authority Title: *Report on Tax Authorities' Childcare Allowance (Kinderopvangtoeslag) Investigation* from: https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf , Page. 26 , Date: July 7, 2023.

13. Kleinnijenhuis, J. (2021, January 7). *OM: Geen strafrechtelijk onderzoek naar Belastingdienst in verband met toeslagenaffaire*. Trouw, 7. Retrieved June 10, 2023, from trouw.nl/politiek/om-geen-strafrechtelijk-onderzoek-naar-belastingdienst-in-verband-met-toeslagenaffaire.

14. *Citizen service number (BSN)*. (n.d.). Netherlands Worldwide. Retrieved July 19, 2023, from https://www.netherlandsworldwide.nl/bsn.

15. Letter to parliament on the analysis of the risk classification model Toeslagen, 21 april 2022, Appendix 4.

16. *Het systeem Fraude Signalering Voorziening (FSV)*. (n.d.). Belastingdienst. Retrieved July 9, 2023, from https://www.belastingdienst.nl/wps/wcm/connect/nl/contact/content/het-systeem-fraude-signalering-voorziening-fsv.

17. Amnesty International, *"Observations to the United Nations Committee on the Elimination of Racial Discrimination* No. 36 on Preventing and Combating Racial Discrimination", June 2019, amnesty.org/download/Documents/IOR4006242019ENGLISH.pdf.

18. "*Contact and confidence: revisiting the impact of public encounters with the police"*, 18 March 2009, Policing and Society, Volume 19, Issue 1, doi.org/10.1080/10439460802457594.

19. *Een ongekende heksenjacht*. (2019, July 8). RTL Nieuws. Retrieved July 12, 2023, from https://www.rtlnieuws.nl/columns/column/4773721/menno-snel-belastingdienst-toeslagenaffaire-ministerie-van-financien.

20. *Dutch childcare benefits scandal*. (n.d.). Wikipedia. Retrieved June 18, 2023, from https://en.wikipedia.org/wiki/Dutch_childcare_benefits_scandal.

21. Geerdink, F. (2021, October 14). *'I cry a lot, every day': victims of the Dutch child benefits scandal fight for compensation*. The Conversationalist. Retrieved June 18, 2023, from https://conversationalist.org/2021/10/14/i-cry-a-lot-every-day-victims-of-the-dutch-child-benefits-scandal-fight-for-compensation/.

22. Advisory committee on implementation of benefits, Omzien in verwondering: Interim-advies, 14 November 2019, p. 48, tweedekamer. nl/kamerstukken/amendementen/detail?id=2019Z22146&did=2019D46007.

23. Council for the Judiciary, “*Algoritmes in de rechtspraak: Wat artificiële intelligentie kan betekenen voor de rechtspraak”*, 2019, rechtspraak.nl/SiteCollectionDocuments/rechtstreeks-2019-02.pdf.

24. Dutch Data Protection Authority Title: *Report on Tax Authorities' Childcare Allowance (Kinderopvangtoeslag) Investigation* from: https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf , Page. 8 , Date: July 7, 2023.

25. Nationale Ombudsman, *Geen powerplay maar fair play*: onevenredig harde aanpak van 232 gezinnen met kinderopvangtoeslag, 9 August 2017, p. 4, nationaleombudsman.nl/system/files/onderzoek/Rapport%202017-095%20Geen%20powerplay%20maar%20 fair%20play_0.pdf .

26. Netherlands Institute for Human Rights, “*Onderzoek College naar tientallen klachten over de Belastingdienst*”, 17 June 2021, mensenrechten.nl/nl/nieuws/onderzoek-college-naar-tientallen-klachten-over-de-belastingdienst.

27. Dutch Data Protection Authority, “*Presentatie onderzoeksrapport ‘De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag’ door Aleid Wolfsen* op 17 juli 2020”, 17 July 2020, autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/toespraak_aleid_wolfsen_onderzoek_kinderopvangtoeslag.pdf.

28. Dutch Data Protection Authority Title: *Report on Tax Authorities' Childcare Allowance (Kinderopvangtoeslag) Investigation* from: https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf , Page. 35 , Date: July 7, 2023.

29. Netherlands House of Representatives, “*Debat over de verklaring van de minister-president en over het verslag van de ondervragingscommissie Kinderopvangtoeslag*”, 19 January 2021, tweedekamer.nl/debat_en_vergadering/plenaire_vergaderingen/details/activiteit?id=2021A00393.

30. Under international human rights standards, the notion of access to justice is enshrined in the European Convention for the Protection of Human Rights and Fundamental Freedoms, Article 6 and 13; EU Charter of Fundamental Rights, Article 47. These rights are also provided for in other international instruments, such as UN International Covenant on Civil and Political Rights, Articles 2(3) and 14; UN Universal Declaration of Human Rights Articles 8 and 10; International Covenant on Economic, Social and Cultural Rights, Article 2; International Convention on the Elimination of All Forms of Racial Discrimination, Article 6

31. AI Now, *AI Now 2017 Report*, 2017, p. 30, ainowinstitute.org/AI_Now_2017_Report.pdf

32. NOS, “*Advocaat: toeslagenaffaire doet denken aan Kafka, liep tegen muur aan”*, 16 November 2020, Retrieved July 9, 2023, nos.nl/artikel/2356771-advocaat-toeslagenaffaire-doet-denken-aan-kafka-liep-tegen-muur-aan.

33. Jan Kleinnijenhuis, "*Gedupeerde ouders willen inzage in hun dossier, maar de Belastingdienst voelt daar weinig voor"*, Trouw, 25 August 2019,Retrieved July 1, 2023, trouw.nl/nieuws/gedupeerde-ouders-willen-inzage-in-hun-dossier-maar-de-belastingdienst-voelt-daar-weinigvoor~b6771696.

34. Dutch Data Protection Authority Title: *Report on Tax Authorities' Childcare Allowance (Kinderopvangtoeslag) Investigation* from: https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf , Page. 77 , Date: July 7, 2023.

35. Prime Minister of the Netherlands, Minister of General Affairs, Letter to Parliament in *response to the report "Ongekend Onrecht"*, 15 January 2021, pp. 2-3, rijksoverheid.nl/documenten/kamerstukken/2021/01/15/kamerbrief-met-reactie-kabinet-op-rapport-ongekendonrecht

36. Chamaki, F. (n.d.). *Automated decision-making impacting society | Knowledge for policy*. Knowledge for policy. Retrieved August 1, 2023, from https://knowledge4policy.ec.europa.eu/foresight/automated-decision-making-impacting-society_en

37. *What is automated individual decision-making and profiling?* (n.d.). ICO. Retrieved August 1, 2023, from https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/automated-decision-making-and-profiling/what-is-automated-individual-decision-making-and-profiling/

38. Demková, S. (2021, November 19). *The Decisional Value of Information in (Semi-)automated Decision-making, by Simona Demková* –. REALaw blog. Retrieved August 1, 2023, from https://realaw.blog/2021/11/19/the-decisional-value-of-information-in-semi-automated-decision-making-by-simona-demkova/

39. Saar Alon-Barkat , Madalina Busuioc, *Human–AI Interactions in Public Sector Decision Making: "Automation Bias" and "Selective Adherence" to Algorithmic Advice*, *Journal of Public Administration Research and Theory*, Volume 33, Issue 1, January 2023, Pages 153–169, https://doi.org/10.1093/jopart/muac007.

40. *Netherlands uncovers $120m 'Bulgarian fraud' benefits scam*. (2013, June 25). BBC. Retrieved August 3, 2023, from https://www.bbc.com/news/av/world-europe-23043543

41. *Scientific Models*. (n.d.). Texas Gateway. Retrieved August 4, 2023, from https://www.texasgateway.org/resource/scientific-models.

42. *Overheid.nl, The Algorithm Register*. (n.d.). Overheid.nl, The Algorithm Register. Retrieved August 4, 2023, from https://algoritmes.overheid.nl/en.

43. *What is an Activity Diagram?* (n.d.). MindManager. Retrieved August 4, 2023, from https://www.mindmanager.com/en/features/activity-diagram/.

44. Ventura, F., Cerquitelli, T., & Giacalone, F. (2018). Black-Box Model Explained Through an Assessment of Its Interpretable Features. Lecture Notes in Computer Science, 138–149.

45. Fong, R., & Vedaldi, A. (2017). Interpretable Explanations of Black Boxes by Meaningful Perturbation. In 2017 IEEE International Conference on Computer Vision (ICCV), 3429–3437.

46. Liu, S., & Vicente, L. N. (2022). Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach. Computational Management Science.

47. Schinckus, C., Gasparin, M., & Green, W. (2022). Opening the black boxes: financial algorithms and multi-paradigmatic research in information technology. Journal of Systems and Information Technology.

48. Google Developers. (2019). Machine Learning Crash Course. https://developers.google.com/machine-learning/crash-course/.

49. Reese, P. (n.d.). *1 Bias, Precision and Accuracy | Download Scientific Diagram*. ResearchGate. Retrieved August 6, 2023, from https://www.researchgate.net/figure/Bias-Precision-and-Accuracy_fig2_305767261.

50. Reese, P. B. (2016). Calibration in Regulated Industries: Federal Agency Use of ANSI Z540.3 and ISO 17025. Proceedings of the 2016 Annual Conference on World Standard Cooperation (WSC).

51. *Siris böse Schwester. Wenn der niederländische öffentliche Dienst Ihre Daten stiehlt – Digital Society Blog*. (n.d.). Alexander von Humboldt Institut für Internet und Gesellschaft | HIIG. Retrieved August 6, 2023, from https://www.hiig.de/siris-boese-schwester-wenn-der-niederlaendische-oeffentliche-dienst-ihre-daten-stiehlt/

52. Altmann, G. (n.d.). *Article 8 - Protection of personal data*. European Union Agency for Fundamental Rights. Retrieved August 6, 2023, from http://fra.europa.eu/en/eu-charter/article/8-protection-personal-data

53. *Article 8: Respect for your private and family life*. (2021, June 24). Equality and Human Rights Commission. Retrieved August 6, 2023, from https://www.equalityhumanrights.com/en/human-rights-act/article-8-respect-your-private-and-family-life

54. Braun, I. (2018, July 4). *High-Risk Citizens*. AlgorithmWatch. Retrieved August 7, 2023, from https://algorithmwatch.org/en/high-risk-citizens/

55. *WaterProof*. (n.d.). Nederlandse Arbeidsinspectie: Home. Retrieved August 7, 2023, from https://www.nlarbeidsinspectie.nl/?utm_campaign=inspectorateszw.nl&utm_source=inspectorateszw.nl&utm_medium=redirect

56. *1.3 Population: demographic situation, languages and religions*. (2022, December 19). Eurydice. Retrieved August 7, 2023, from https://eurydice.eacea.ec.europa.eu/national-education-systems/netherlands/population-demographic-situation-languages-and-religions

57. *Expert Maths Tutoring in the UK - Boost Your Scores with Cuemath*. (n.d.). Expert Maths Tutoring in the UK - Boost Your Scores with Cuemath. Retrieved August 7, 2023, from https://www.cuemath.com/sample-size-formula/

58. Coglianese, C., Shapiro, S., Katz, J., & Rodrigues, J. (2022, January 10). *Empathy in an Automated State*. The Regulatory Review. Retrieved August 7, 2023, from https://www.theregreview.org/2022/01/10/coglianese-empathy-automated-state/

59. Kuziemski M, Misuraca G. AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. Telecomm Policy. 2020 Jul;44(6):101976. doi: 10.1016/j.telpol.2020.101976. Epub 2020 Apr 17. PMID: 32313360; PMCID: PMC7164913.

60. Ramya Srinivasan, Beatriz San Miguel González,The role of empathy for artificial intelligence accountability, Journal of Responsible Technology, Volume 9, 2022, 100021, ISSN 2666-6596, https://doi.org/10.1016/j.jrt.2021.100021. (https://www.sciencedirect.com/science/article/pii/S2666659621000147)

61. *Art. 22 GDPR – Automated individual decision-making, including profiling - General Data Protection Regulation*. (n.d.). GDPR. Retrieved August 7, 2023, from https://gdpr-info.eu/art-22-gdpr/

62. Kate Goddard and others, “Automation bias: a systemic review of frequency, effect mediators, and mitigators”, Journal of the American Medical Informatics Association, Volume 19, Issue 1, 16 June 2011, https://academic.oup.com/jamia/article/19/1/121/732254.

63. Price WN. Big data and black-box medical algorithms. Sci Transl Med. 2018 Dec 12;10(471):eaao5333. doi: 10.1126/scitranslmed.aao5333. PMID: 30541791; PMCID: PMC6345162.

64. *Accuracy and precision*. (n.d.). Wikipedia. Retrieved August 8, 2023, from https://en.wikipedia.org/wiki/Accuracy_and_precision

65. Gillis, A. S. (n.d.). *What is Machine Learning Bias? | Definition from WhatIs*. TechTarget. Retrieved August 8, 2023, from https://www.techtarget.com/searchenterpriseai/definition/machine-learning-bias-algorithm-bias-or-AI-bias

66. Rudin, C., & Radin, J. (2019). Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson From an Explainable AI Competition. Harvard Data Science Review, 1(2). https://doi.org/10.1162/99608f92.5a8a3a3d

67. Benefits agency UWV breached privacy rules with IP monitoring. (2023, July 15). *DutchNews.nl*. https://www.dutchnews.nl/2023/07/benefits-agency-uwv-breached-privacy-rules-with-ip-monitoring/

68. Chapter European Union Fosters Ethical AI in the Public Administration , New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence, 2023, Volume 1430, ISBN : 978-3-031-14858-3, António Costa Alexandre, Luís Moniz Pereira

69. “What Is Machine Learning? | IBM,” n.d., https://www.ibm.com/topics/machine-learning.

Appendix:

The interview was conducted over the phone on July 19, 2020, and was conducted in Arabic. Below is the transcript of the interview translated into English.

Interviewer: How old are you?
Interviewee: 38 years old.

Interviewer: What is your current salary?
Interviewee: My current salary is 20,000 per year. This year's salary is about 20,000, whereas last year's salary was about 35,000.

Interviewer: How many children do you have?
Interviewee: 4 children

Interviewer: What is your nationality?
Interviewee: Syrian, but I have Dutch nationality also.

Interviewer: So you have Dutch nationality?
Interviewee: Yes, I obtained it about 2 or 3 years ago.

Interviewer: Did you have Dutch nationality when the problem occurred?
Interviewee: No

Interviewer: When did you arrive in the Netherlands?
Interviewee: 2015

Interviewer: May I ask whether you were married or single when the problem occurred?
Interviewee: Yes, I was married, and I have 4 children.

Interviewer: What is the approximate distance between your house and the kindergarten?
Interviewee: The kindergarten is about 2 km away from my work and approximately 10km from my house.

Interviewer: What is the typical or average number of hours that the children spent at the kindergarten each day before the corona pandemic?
Interviewee: Back then, I had 3 children who attended the after-school kindergarten program called "na school blijven," which is different from regular "kinderopvang." They did not go every day; usually, they attended one or two days a week, spending about 6 hours there. I used to pick them up directly after finishing my work, and I never left them until 18:00. My work hours were from 9 in the morning until 1 in the afternoon, so my son had to use both periods, from 7 am until 12 pm and from 12 pm until 6 pm. However, I didn't leave him until 18:00 because he didn't feel comfortable with the people in the kindergarten. He used to cry a lot, and I could tell that the kindergarten wasn't suitable for him. Whenever I passed by the kindergarten on my

bicycle while going from one client to another, I would see him standing outside and crying. So, I waited until I finished my work to bring him home.

Interviewer: The kindergarten fees are paid partly by the tax authorities (belasting) and partly by you. How much do you pay, and how much does the tax authority pay?
Interviewee: We pay about 100 to 150 euros, and the tax authority pays about 2700 euros.

Interviewee: Can you please provide information about the organisations that contribute to payments for the kindergarten program?
Interviewee: skip

Interviewer: Do you have any other debts with the Dutch government?
Interviewee: No no.

Interviewer: Have you experienced any other mistakes with taxes?

Interviewee: No, I haven't. I have an accountant because when I first came to the Netherlands, I applied for a family reunion. Initially, I came with my daughter. You know that when you are with your daughter only, you receive the salary of a single mother. However, when my husband and my other children came to the Netherlands, the tax authorities said that they had given me the salary of a single mother, even though I was not a single mother. Consequently, they asked me to return the money, and I promptly did so. Since my husband arrived, I have been consistently paying taxes without interruption. This has been the case from 2017 until now, and I make monthly payments. If you were to see my bank account, you would notice that I've been repaying every month. Frankly, it is exhausting to deal with them. They asked for the money for the single mother issue, covering the period from 2015 to 2017. Then, in 2019, after the corona pandemic, I faced the kindergarten issue, but this time, it involved a significant amount of money. I understand that we are working and not using social services, but 8000 euros is not an easy amount to pay. It would be much better to spend this money on my family. The situation is quite frustrating because it was their mistake to give me money and then ask for it back, even though I didn't request it. It would have been more sensible to allocate the funds directly to the kindergarten from the beginning.

Interviewer: What type of allowance is given? Is it a rent allowance or a child allowance?
Interviewee: skip

Interviewer: When did you receive the letter of the debt?
Interviewee: I received the notice regarding the single mother issue in February 2017, precisely. The resolution for the refund of the kindergarten money was in June 2020, and I received the

notification in August 2020. When I received the letter I couldn't believe the contents of the letter. I knew that it was just a big mistake that will get resolved once I call the Belastingdients.

Interviewer: What does the letter say?
Interviewee: skip

Interviewer: What is the amount of the debt?
Interviewee: The single mother debt was about 3000 or 4000 euros. I don't remember exactly because the issue dates back to 2017, and it's been quite a while.

Interviewer: Did the financial aid stop directly after the letter?
Interviewee: Honestly, we did not receive any aid for a certain time because both my husband and I were working. However, I had to get the aids later on because I became handicapped and could not work anymore. So they started giving me the aids again, and currently, I am receiving them.

Interviewer: I mean, did they immediately stop the financial aid after sending you the debt post and asked you to return the money, or were they still providing you with financial aid?
Interviewee: No, it was an ongoing process of receiving and returning the money. They pay, and I return it. I have to make payments on every 21st of the month, and I receive the money on the 20th. So, when I receive the money, I immediately return it to them.

Interviewer: I mean, did they directly stop the kindergarten aid after sending you the post?
Interviewee: No, the kindergarten aid was actually stopped before they sent the post. To be more exact, the kindergarten aid was stopped in November 2019, and I received the post for payment in August 2020. I mean the post with the final decision to pay.

Interviewer: So, they stop it after they send you the post?
Interviewee: No, they stop it first, and then they send the post.

Interviewer: For how many months did the aid stop?
Interviewee: skip

Interviewer: How was your social and psychological status during that time with your husband and family? You totally have the freedom to answer this question or not.

Interviewee: To be honest, when we faced the reality that our only option was to pay the tax authorities the full amount of 8000 Euro, our lives turned upside down and deteriorated drastically over time. We were under significant financial distress, and it had a severe impact on our psychological well-being. We were extremely anxious and shocked by the burden of dealing with an 8000 euro debt. My husband blamed me, but I don't believe it was entirely my fault. I handle the paperwork and payments at home, while my husband focuses on working and earning money. He relies on me for managing such matters. He insisted that I should have stopped the aids from the beginning, but honestly, I was unaware of the legal procedures like "uitvragen" and "aanvragen." I didn't know what to do in this situation. This created a problem between us, and the shock of the situation was overwhelming. The stress and anxiety took a toll on my health. I developed a neurological disease due to the stress, as diagnosed by the doctors. I spent one and a half years of my life going from one hospital to another and from a psychologist to another. They determined that I have a neurological issue, likely caused by the immense strain this situation put on me.

Interviewer: When communicating with the tax authorities, did they provide a clear reason and explanation for stopping the financial aid and claiming the refund of the required amount?
Interviewee: No, and honestly, the employee was very disrespectful in his manner of speaking. He literally said, "mevrouw, the money was transferred to your account, and you have spent it." I told him that I haven't spent the money, but he insisted that he didn't know where I spent it, whether on shopping or something else. I reiterated that I gave the money to the kindergarten, but he claimed that the kindergarten returned the money. I told him that they only returned 10 euros, and if he wanted that back, he could have it, but I couldn't afford to pay 8000 euros. He said he would check the situation and send a reply in a post in 4 weeks. This particular employee was essentially the reason for the problem. He mentioned that I could make an objection after receiving the post, but I never received any post, and therefore, I didn't submit an objection. After a month, I called the belasting back and explained the situation, mentioning that I was waiting for the objection form to make an objection. I asked if I could at least pay less than 8000 euros since I couldn't afford such a huge amount. The person I spoke to denied that I had asked for the objection form and claimed that I was already in the payment due phase, which meant I couldn't make an objection anymore.

Interviewer: Was it disclosed that an algorithm has been used?
Interviewee: to be honest I don’t know what does the word algorithm means
Asiea reply

Interviewer: Was there any cooperation from the tax authorities' side?
Interviewee: skip

Interviewer: Were you able to file an objection, and did that help reduce the amount of the debt?
Interviewee: skip

Interviewer: Have you appointed a defense attorney?
Interviewee: No, everyone advised me against doing so. It's generally difficult to get your rights back and win a case against the belasting (tax authorities). They are heavily protected by the government, and it's unlikely to work in our favor. Even my husband's Company Accountant said the same thing. He advised that although I could provide all the evidence and papers, winning the case would be challenging.

There was a verdict signed by the king stating that families with such problems would receive compensation of 30,000 euros. I went to an accountant in Almere or Lelystad and provided the necessary papers to apply for the compensation. He said I would receive 40,000 euros as compensation and assisted me in completing the compensation application for a fee of 150 euros. However, after some time, I received a post stating that the compensation verdict had not been implemented, and even if it were, I wouldn't be eligible to receive it.

Interviewer: What was the final verdict?
Interviewee: skip

Interviewer: Do you think it is possible to avoid the error that occurred by providing the tax site in other languages?
Interviewee: Of course, of course. If I had known from the beginning what I am reading without using Google translation, because Google translation is really bad. So, if I am able to read exactly what is written without Google translation, I would not have made any mistakes, and I would do everything correctly one hundred percent. But using Google Translate and the poor language skills that we have learned here, I could not understand anything. You click "yes, yes, yes," and in the end, you get a debt in return.