Who Oversees Artificial Intelligence Systems: A Descriptive Analysis of Ethical Concerns over the Dutch Childcare Benefit Scandal Utilising the Networked Systems Ethics Guidelines

Francisco Pereira, Gijs Gubbels, Dragos Pop, Lan Chu, and Robin van den Berg

University of Amsterdam

Word count: 2020

1 Introduction

Data science-based decision making has become increasingly important over the last couple of years to increase the efficiency of government institutions. Although handing over decision making from employees to algorithms seems to be a good method for preventing bias, it has become clear that machine learning algorithms are not immune to biases either [13]. There are great concerns about the use of biased data and a case exhibiting this was the Dutch childcare benefits (CB) scandal that led to the resignation of the Dutch government on the 15^{th} of January 2021. CB was introduced to the Dutch social welfare system in 2004. In 2013, the Netherlands uncovered a €10 million "Bulgarian fraud" surrounding childcare allowance [11]. In response to this, a "Tackling Fraud" Committee was established to develop an anti-fraud strategy at a large scale. The Tax Administration (TA) took a strict approach and used an automated algorithm to detect applications with a high potential of fraud. This childcare benefit scandal came to public attention when news outlets reported the algorithm wrongfully accused thousands of families of fraud in September 2018 [10]. Here, the classification algorithm designed to detect fraudulent behaviour falsely labelled 26000 parents as fraudsters [5], with an imbalance towards parents with an immigration background, forcing the victims to pay back large sums of money that were rightfully received.

In this paper, an analysis will be done of the Dutch CB scandal as it shows the dangers of ML-based decision making and can teach us how to avoid possible pitfalls. We assess ethics aspects applying a subset of the Networked Systems Ethics (NSE) guidelines applicable to this case study. For this article, we will assess the technology design, review the power relations between and the risks for the parties involved.

Firstly, an objective analysis of the different aspects of the case will be given, focusing on the aspects that pertain to the terms mentioned above. Subsequently, the case will be discussed based on the analysis, supplying the viewpoints of the authors and suggestions for improvement. Lastly, the article will be concluded with a summary of our main observations and discussion points regarding the case study.

2 Analysis

2.1 Data Integration, Technology Design and ethical concerns

Data integration and dissemination

The data used for the fraud detection is acquired from the *Gemeentelijk Basisregistratie*

(GBA). This database includes information about one's name, address, place of residence, marital status, nationality, and dual nationality if applicable. After January 6^{th}, 2014, the GBA was succeeded by *Basis Registratie Personen* (BRP) [14]. An important difference between GBA and BRP is that the latter does not include dual nationality anymore. *Beheer van Relaties* (BVR), which is interfaced with GBA and BRP, is the "umbrella" system used by TA to access the data from the population administration system through over-night processing. The downstream application *Toeslagen Verstrekking Systeem* (TVS) is used for receiving allowance applications, paying out and refunding of the child benefits, and monitoring the implementation [14].

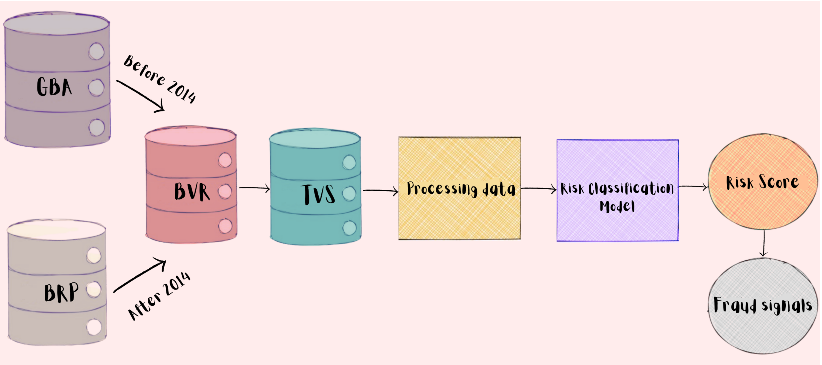

Fig. 1: Architecture design of Risk Classification model

The use of a risk classification model for fraud detection

The TA utilises a risk classification model which uses a dozen of (after-processing) indicators from TVS since 2013, outputting a risk score between 0 (minimum) and 1 (maximum) for an applicant. This is a self-learning model that is trained with examples of correct and incorrect applications. The outputted risk scores are used to indicate applications with a high potential of fraud and request further assessment accordingly. [14]

Dutch/non-Dutch nationality indicator in risk model and fraud detection

To be eligible for benefits, the legal condition is that the applicant must have a valid residence permit. As stated by TA, due to some technical and practical constraints, this information was not accessible and the nationality has been used as a substitute to derive a Dutch/non-Dutch indicator for the risk model [14]. After receiving a fraud signal, the employees of the Benefits department requested the first and second nationality of the detected applicants, which were used when manually assessing applications. The Dutch Data Protection Authority (AP) later determined the use of nationality unnecessary for the risk classification model as well as for signalling organised frauds [14].

The undesired ethical effects

First, when it comes to the storage design of the system, there are some limitations. Under GDPR, data that is no longer supposed to be accessible should be removed from the system. Dual nationality was wrongfully preserved and used in TVS (which was fed from GBA)[14]. TVS system should have been designed in a way that ensured

dual nationality would no longer be accessible. This concerns the ethical issues in data dissemination and storage as part of data governance.
Secondly, the “self-learning” model used nationality data as a predictor for the risk score [4]. If the data used to learn the model reflects existing social biases, the algorithm is likely to incorporate these biases. This allowed the model to assign predictive importance to the nationality feature. Consequently, the model negatively discriminated against people with other nationalities than Dutch, assigning higher risk scores, as confirmed by the Director of Benefits Department [14].

2.2 Imbalanced positions for involved parties

The relations of the stakeholders are important in evaluating the current case. We deal with two parties, namely the Dutch TA and the recipients of the CB [6]. The power relation is evident, as the former is an institution of the government, while the latter are people from low-income families that require CB they can only get from the former. In addition, of these two groups, the TA are the only ones that can access the model and the information they have collected [9]. The prospective recipient files the request for the CB and provides the TA with the needed information. After this, however, all procedures are controlled by the opposite party. Furthermore, a technique as applied by the TA relieves them from explaining the selection procedure and, therefore, redirects responsibility towards an algorithm. The choices made by the algorithm are opaque for the user and consequently also for the citizens [9]. Since the inner workings of the methods used for fraud detection are not known to the accused, it makes it hard to fight the choices made by the TA. The current classification method, therefore, increases the power gap between the two stakeholders.

Looking at the possible gain and loss for both parties, we also find a strong imbalance. With regards to the benefits of the model, we find that a boost in efficiency is possible. Such a boost implies less money is required which will benefit the society since the money can be spent in other places. Additionally, the project should have led to a higher accuracy of fraud detection, leading to fewer allowances paid out. With the current outcome, however, fines had to be paid to the wrongfully accused families, leading to higher costs for society as the fines were paid by the government and therefore the taxpayer [1]. The main disadvantages, however, lie with the families receiving CB. These families are subjected to discrimination and subsequently required to pay back money that was rightfully received. Often not in the position to find legal aid, they were also not able to object to the claims of the TA. In addition, several hundred parents were incorrectly denied the pay-out suggesting they did not request CB [8]. Here, we see a great power imbalance as the accused are powerless against the verdict of the TA.

2.3 Unforeseen Consequences

The government’s role is to instil order, protect citizens and provide goods/services that help individuals live in society. However, the government is vulnerable to fraudulent individuals that aim to deceive the system. As mentioned before, data-driven algorithms are used to tackle this type of situation but they can be double-edged swords. On one side benefits arise (e.g. efficiency, reduced costs, etc.), on the other, historical biases and discrimination embedded into data sets may quietly penetrate decision making [3]. The latter was the case. Serious harm was caused to families that were accused of such crimes and the political stability of the country suffered, as Rutte III cabinet resigned [6]. This situation resulted from impetuous system implementation, as the information on applicants nationality was not supposed to be accessible since 2014.

It is of high priority to be aware of the consequences that arise from the outcomes of such systems. Identifying someone as a fraudster is a serious claim and consequences vary between incarceration, probation, fines and restitution [7]. Wrongfully accusing someone of fraud has a tremendous impact on the individuals. In this case, many were driven to financial ruin as they were asked to repay the state large sums of money, which tended to initiate a downward spiral of events such as depression, divorces and many people suffered discrimination as they were seen as fraudsters by society [2]. Given the impact of false-positive accusations in this scenario, one of the main priorities when implementing the data and AI system, should have been avoiding it, which appears to have been neglected.

The current Dutch government has the political belief that outsourcing and automating certain government processes results in higher effectiveness and efficiency [13]. Outsourcing and automation mean less control and less attention to individual cases. However, less control and attention to individual cases also means a higher chance of errors. In this scenario, errors are not allowed since it questions the government's role (as first defined in this section) and threatens political stability, ultimately impacting millions of people. The government should have been notified by the TA of the possible risks of the system which remains unclear if it was the case.

3 Discussion

AI's bias' caveats and precarious implementation

The main point this paper tries to convey is the socially harmful effect biased data and unsatisfactory implementation of AI can have on its subjects. Essentially, using a sensitive feature in training the algorithm made it prone to biased outcomes. Not only this practice is unethical due to the associated risk of discrimination, but it is also illegal since it violates the GDPR, as well as conceptually erroneous, considering that nationality is not relevant when addressing fraud detection. The best solution, in this case, would be eliminating the sensitive feature and the correlated variables altogether, so that the former one cannot be later derived from another feature. Alternatively, ensuring demographic parity between the Dutch nationals and the non-Dutch, taking the nationality as the protected class, would also combat discrimination because the two groups would have the same probability of being classified as fraudsters. As an alternative, applying equalised odds would also increase the fairness of the algorithm, especially when considering the harmful effects of false positives.

Concerning the ML model's implementation, before deployment, the Dutch TA should have ensured the model's precision is high and, most importantly, the number of false positives is very low because, as it was argued in the previous section, it has drastic consequences on the innocent individuals that were classified as fraudsters. Moreover, the explainability of the model should be addressed and improved. Transparency plays an important role for the TA, as it allows the investigators to understand the reason for the high risk score (which they currently did not get) and points out the reason behind the decision of classifying applicants as fraudsters. With a proper explanation, the accused would be able to better plead their innocence. However, not only did the algorithm fail to correctly select fraudulent cases, but manual research also failed to accurately identify the true positives, suggesting carelessness from the researchers' side.

4 Conclusion

In this article, we discussed the case of the CB scandal that occurred in the Netherlands. The TA has made fundamental ethical missteps in assessing the CB applications. It was concluded that there was improper use of nationality and dual nationality data, which should not even have been accessible, thus illegal and careless. Furthermore, we found that a strong power imbalance was present between the two stakeholders, making the false positives that occurred significantly more impactful, as they were often not remedied. Lastly, the false positives also had severe results for the families, often driving them into severe financial problems. To prevent subsequent similar cases, it is imperative to analyse biased ML-based decision making, especially as the Netherlands seems to be at the forefront of predictive policing, which has the potential to lead to catastrophic consequences [12].

References

1. 70.000 gedupeerde kinderen toeslagenaffaire krijgen tegemoetkoming tot 7500 euro. RTLnieuws https://www.rtlnieuws.nl/nieuws/politiek/artikel/5237810/geld-voor-70000-gedupeerde-kinderen-toeslagenaffaire
2. The childcare benefits scandal: voices of the victims https://www.dutchnews.nl/news/2021/01/the-childcare-benefits-scandal-voices-of-the-victims/
3. Data-driven discrimination: a new challenge for civil society https://blogs.lse.ac.uk/impactofsocialsciences/2018/07/10/data-driven-discrimination-a-new-challenge-for-civil-society/
4. The dutch benefits scandal: a cautionary tale for algorithmic enforcement. EU Law Enforcement https://eulawenforcement.com/?p=7941
5. Dutch government resigns after childcare benefits scandal. CNBC https://www.cnbc.com/2021/01/15/dutch-government-resigns-after-childcare-benefits-scandal-.html
6. Dutch government resigns after childcare benefits scandal. The Guardian https://www.theguardian.com/world/2021/jan/14/dutch-government-faces-collapse-over-child-benefits-scandal
7. Dutch penalties for fraud https://www.government.nl/latest/news/2020/12/03/online-fraud-with-new-techniques-to-be-more-severely-punished
8. Honderden toeslagenouders onterecht tegemoetkoming geweigerd. Trouw https://www.trouw.nl/politiek/honderden-toeslagenouders-onterecht-tegemoetkoming-geweigerd~bf41517b/
9. Methods used by dutch tax and customs administration unlawful and discriminatory. Autoriteit Persoonsgegevens https://autoriteitpersoonsgegevens.nl/en/news/methods-used-dutch-tax-and-customs-administration-unlawful-and-discriminatory
10. Tax authorities deliberately worked against parents who were entitled to childcare allowance https://www.trouw.nl/nieuws/belastingdienst-werkte-ouders-die-recht-hadden-op-kinderopvangtoeslag-bewust-tegen~bf13daf9/
11. Up to 4 years in prison for bulgarian benefits fraud https://nltimes.nl/2015/05/19/4-years-prison-bulgarian-benefits-fraud
12. We sense trouble: Automated discrimination and mass surveillance in predictive policing in the netherlands. Amnesty International https://www.theguardian.com/world/2021/jan/14/dutch-government-faces-collapse-over-child-benefits-scandal
13. Cath, C., Jansen, F.: Dutch comfort: The limits of ai governance through municipal registers. arXiv preprint arXiv:2109.02944 (2021)
14. Persoonsgegevens, A.: De verwerking van de nation-aliteit van aanvragers van kinderopvangtoeslag, https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/onderzoek_belastingdienst_kinderopvangtoeslag.pdf