MASTER THESIS

Towards the design of legally privacy-proof and ethically justified data-driven fraud risk assessment algorithms

Xadya van Bruxvoort

PROGRAM
Master of Science Business Information Technology

SPECIALISATION
Data Science & Business

FACULTY OF ELECTRICAL ENGINEERING, MATHEMATICS AND COMPUTER SCIENCE

EXAMINATION COMMITTEE
dr. ir. Maurice van Keulen
mr. dr. Lesley Broos
dr. ir. Jeroen van Oostrum
Pim de Waard MSc.
Thomas Stolker MSc.

DOCUMENT CONFIDENTIALITY
Public

22 March 2021

M&I /Partners/

Amersfoort

UNIVERSITY OF TWENTE.

Abstract

A growing amount of personal data are being generated every day and municipalities are starting to see the value of these data for fraud detection. Municipalities possess few resources to properly assess social security assistance clients for fraud. This offers potential for algorithms to help them better fulfill their lawful duties of ensuring social benefits assistance are granted justly. Much social security assistance fraud is currently undetected, leading to an estimated 1 billion euros in social damages in the Netherlands annually. However, recent applications of algorithms in fraud detection have led to negative reactions in the media, based on ethical and legal objections. Municipalities therefore struggle with public communication on the topic, in fear of resistance from the public. Nonetheless, the potential benefits of algorithms could allow law enforcers to broaden their scope of potential fraud cases by covering unknown blind spots and improvements in objectiveness. Another benefit could be to detect fraud cases in an earlier stage, thus helping customers earlier on. The Dutch municipality Amersfoort (Gemeente Amersfoort) is an example of a party interested in developing and applying a fraud-detection algorithm, but is also in need of concrete knowledge on how to approach this from a legal and ethical perspective. To simplify this challenge, frameworks need to be designed to provide proper guidelines on the legal and ethical considerations during the design and implementation phases of such a fraud detection algorithm. This study aimed to research, design and validate such frameworks.

Conventionally, different sources can produce signals for social security fraud. A traffic light model may decide whether these signals are further investigated before resulting in a verdict on the suspected fraud. To compare law enforcers with the way algorithms behave for fraud detection, a random forest model was trained and tested on a dataset from Gemeente Amersfoort and compared with interviewed law enforcers. The latter were found to look at different data tables than algorithms. Compared to algorithms law enforcers create social context for data subjects and tend to empathize with them more. However, algorithms are able to process many more customers for fraud detection. Legal guidelines were designed based on national laws, regulations and relevant court cases. These guidelines were combined into a legal framework, which was validated by an expert interview with a lawyer. An ethical framework was created by combining multiple existing frameworks and conducting a brainstorm session, with the main focus points on beneficence, non-maleficence, autonomy, justice and explicability. This framework was then validated by applying it to two case studies and comparing its outcomes to the opinion of the general public. To assess whether the inclusion of conventional data attributes are legitimate, accurate, justified and effective, attribute selection was applied on the dataset to remove non-predicting, correlated, and ethically questionable data. A timeline was created of which parts to consider in every phase of the algorithm-creation process.

Validating the frameworks proved them to be robust tools for detecting focus points on legal and ethical aspects for fraud detection algorithms by municipalities. The ethical framework looks at the process from a broader point of view; from data to protocols. It was also stressed that ethics are important during the entire process from design phase to after end of life. Where previously concrete steps and information were missing, the frameworks offer clear guidelines for municipalities looking to explore fraud detection using an algorithm.

Nevertheless, the two frameworks were validated using relatively small number of cases. Real-world application in other cases or contexts are needed to further validate the frameworks. From interviewing law enforcers, it was found that human experts look at different data attributes depending on the type of fraud they suspect. Hence, different algorithms for different types of fraud might be interesting, to potentially achieve better accuracy or reduce personal data needed per algorithm. With some alterations, these frameworks could also potentially be converted for implementation of algorithms in a different context. Given the relevance of this topic, other instances have recently released related frameworks and guidelines. Combining these frameworks with the ones designed here might be interesting for a more all-round solution in the near future.

Acknowledgements

This thesis was written during a time where the world as we knew it suddenly changed. Because of this, the time it took for me to write this thesis was doubled, which took a lot of patience. Not only for me, but also for the people around me. I am deeply grateful they granted me their patience and support throughout this entire year. Finishing this thesis also marks the end of my student life at the University of Twente, where I had the honour to be part of many organisations and activities. During my bachelor Creative Technology, I discovered the passion for the combination of business or management and IT, which led me to this Master's degree. In the meantime I have learned invaluable knowledge about myself and about the world at my study association Proto, during a full-time board year and other committee work, and at Student Net Twente, during board years. It has been a great time, but great times are ahead as well, and it's time to move to the next phase.

This thesis could not have been written without a few important people.

Maurice van Keulen was honestly a 'rots in de branding'. I could not have wished for better supervision during both my Bachelor thesis and especially during this wild ride of a Master thesis. Maurice was incredibly helpful on the professional parts of this thesis, but also very thoughtful on personal level. I am going to miss the weekly progress meetings, which were more often than not also there for general chit-chat about the world. I am grateful for all your help Maurice and I wish you all the best in the future.

Furthermore, I would like to thank Jeroen van Oostrum for all his help during the past year. We started with a brainstorm from a very practical perspective and built a thesis around this. He also helped tremendously in the communication aspect of this thesis; my infographics would have looked way worse otherwise. But also on a personal level Jeroen helped me by keeping interest even though I could not do much, which I very much appreciate, and by enlightening every situation with humour. I hope to be working more with Jeroen in the future.

I cannot thank Thomas Stolker enough. Thomas' input was invaluable, and he often thought along with me about questions I had. More importantly, he showed me much of the practice from which I learned a lot. From political questions I never thought about, to stakeholders I never even heard of before. Thomas was also a pleasure to work with, reacting to all my questions quicker than I could have wished, and it was always 'gezellig' to have our biweekly phone calls. I think Gemeente Amersfoort must be very grateful for having you there. Thanks for all the help and the fun we had in the meantime!

Pim de Waard helped this thesis and myself improve a lot by asking critical questions at the right moments. This made me look at things differently, which was very helpful. During my time at the office, before Covid-19 hit, I enjoyed your patience and presence. Let's hope this time will come back and we can joke at the coffee machine.

I would also like to thank Sake Alkema in this part. It is not easy working eight hours a day on the same dining table with two clicky mechanical keyboards. But you helped with everything; bringing coffee when it was desperately needed, letting me use you as rubber duck to bug fix, reading through parts of this report I was unsure about, listening, and all the general support. Thank you.

Lesley Broos helped this thesis by providing interesting feedback from a different perspective and making sure this thesis was involving all aspects needed (also legal aspects). His knowledge is exceptional. Thanks.

I would also like to thank M&I/partners for welcoming me in their organisation like I was one of them from the beginning. The number of people who thought along with this thesis is rather large, and I have not experienced such a welcoming organisation with warm people before. In special I would like to spend a few extra words to thank Anne-Marie for your extra help this summer, but also for being so warm and welcoming. Tobias for the introduction to Gemeente Amersfoort. Patrick, for the infinite knowledge you send to me in my inbox.

Gemeente Amersfoort deserves a huge thank you. They let me use their knowledge, information and data which was the foundation of this report. They also made me feel very welcome, already from our first meeting. Within Gemeente Amersfoort, I would also like to thank a few people in particular. Ate for providing feedback, thinking along and the pleasant meetings. All interviewees at Gemeente Amersfoort, thank you for your time and input, it was very much appreciated.

Last, I would like to thank my parents and friends for the support during my entire student life, and also the last year.

Xadya van Bruxvoort
Arnhem, 22 March 2021

List of Figures

Figure 1 - Reasons for the big brother award of SyRI 3
Figure 2 - DSRM of Peffers et al. (2007) 7
Figure 3 - Location of Gemeente Amersfoort 11
Figure 4 - Difference in AI, machine learning and deep learning according to Ian Goodfellow, Yoshua Bengio & Aaron Courville in Holzinger et al. (2018) 14
Figure 5 - Ethical framework AI4People 18
Figure 6 - Guidelines of van Wynsberghe, Been and van Keulen 19
Figure 7 - Data Ethics Decision Aid (English version of DEDA) 21
Figure 8 - Roadmap for Artificial Intelligence Impact Assessment (AIIA) 22
Figure 9 - Black box programming [source Wikipedia.org] 23
Figure 10 - Fraud detection at Gemeente Amersfoort 30
Figure 11 – Legal guidelines for social security fraud detection 46
Figure 12 - Algorithm in sociotechnical system 51
Figure 13 - Reference date per client, coloured parts are data that are taken into account, grey parts are data that are being discarded 53
Figure 14 - Altering of data set in case of double rows by aggregation 53
Figure 15 - Altering of data set in case of double rows by creating Booleans 54
Figure 16 - Example of a decision tree [source: https://www.jeremyjordan.me/decision-trees/] 55
Figure 17 - Decision tree for information split 57
Figure 18 - Decision tree for Gini index split 58
Figure 19 - Accuracy of the RF algorithm based on the number of predictors 59
Figure 20 - Differences between the law enforcers and the algorithm 72
Figure 21 - Project framework 78
Figure 22 - Framework results SyRI 91
Figure 23 - Framework results Gemeente Amersfoort 96
Figure 24 - Implications of the research on effectiveness, accuracy, legitimacy and just on the designing phase of fraud detection algorithms 114
Figure 25 - Ethical framework for the implementation of fraud algorithms within public organisations. 138

List of Tables

Table 1 - Capital limit social assistance 2020 ____ *12*
Table 2 - Confusion matrix for information split ____ *56*
Table 3 - Confusion matrix for Gini index ____ *57*
Table 4 - Confusion matrix of the RF algorithm ____ *59*
Table 5 - Top 20 most important variables of the RF algorithm ____ *60*
Table 6 - Confusion matrix Lasso algorithm ____ *61*
Table 7 - Confusion matrix Rulefit algorithm with Caret ____ *61*
Table 8 - Attribute usage of Rulefit with Caret ____ *62*
Table 9 - Confusion matrix MARSplines ____ *63*
Table 10 - Most important variables MARSplines ____ *63*
Table 11 - Accuracy of the different algorithms ____ *64*
Table 12 - Overview of tables and variables within those tables used per algorithm ____ *66*
Table 13 - Results of prediction of fraud cases of the law enforcers ____ *70*
Table 14 - Summary of filled in ethical framework on SyRI ____ *90*
Table 15 - Comparison of the framework of SyRI to the court case and the media ____ *93*
Table 16 - Summary of filled in ethical framework on Gemeente Amersfoort ____ *96*
Table 17 - Comparison of the framework on Gemeente Amersfoort to the media and the interviews ____ *99*
Table 18 - Stakeholders of SyRI ____ *147*
Table 19 - Expectation and wishes of stakeholders of SyRI ____ *148*
Table 20 - How the stakeholders of SyRI are affected ____ *149*
Table 21 – Position of stakeholders of Gemeente Amersfoort's algorithm ____ *165*
Table 22 – Wishes of stakeholders of Gemeente Amersfoort's algorithm ____ *166*
Table 23 - How the stakeholders of Gemeente Amersfoort's algorithm are affected ____ *167*

List of Acronyms

AI	Artificial Intelligence
A/IS	Autonomous and Intelligent Systems
AP	Autoriteit Persoonsgegevens
CBS	Centraal Bureau voor Statistiek
DPIA	Data Protection Impact Assessment
FNV	Federatie Nederlandse Vakbeweging
GDPR	General Data Protection Law (nl: *AVG*)
NOS	Nederlandse Omroep Stichting
RF	Random Forest
SyRI	Systeem Risico Indicatie
SUWI	Wet structuur uitvoeringsorganisatie werk en inkomen
UWV	Uitvoeringsinstituut Werknemersverzekeringen

Table of Contents

Abstract ii

Acknowledgements iv

List of Figures vi

List of Tables vii

List of Acronyms viii

Table of Contents ix

1. Introduction 1

1.1 Introduction 2

1.2 Problem statement 3

1.3 Case objective 4

1.4 Research questions and method 5

1.5 Design Method 6

1.6 Practical note 8

1.7 Structure 8

2. Background 9

2.1 Municipalities in the Netherlands 10

2.1.1 General structure 10

2.1.2 Gemeente Amersfoort 11

2.2 Social security 11

2.3 (Fraud detection) Algorithms 12

2.4 Artificial Intelligence 13

2.5 Stakeholders 14

2.6 Background of using algorithm within municipalities 15

2.6.1 Why do municipalities use fraud detection algorithms? 15

2.6.2 How far are municipalities with this? 16

2.6.3 Case example: SyRI 16

2.6.4 Conclusion 17

2.7 Ethical frameworks 17

2.7.1 AI4People 17

2.7.2 Guidelines for using data from online networking sites 18

2.7.3 Partnership on AI 19

2.7.4 IEEE general principles 20

2.7.5 De Ethische Data Assistant (DEDA) 21

2.7.6 Artificial Intelligence Impact Assessment 22

2.7.7 Other ethical frameworks 23
2.8 Explainable AI 23
2.9 Fraud detection algorithm Totta Data Lab 24
3. Fraud Detection Process 26
3.1 Method 27
3.2 Interviews 27
3.2.1 Project manager 28
3.2.2 Law enforcer 28
3.2.3 Team manager 28
3.2.4 Alderman 29
3.3 Results 30
3.3.1 Fraud detection process 31
3.3.2 Background information on the fraud detection process 32
3.3.3 Preventing discrimination 32
3.3.4 Gemeente Amersfoort compared to other municipalities 33
3.3.5 Benefits and drawbacks of a fraud detection algorithm 34
3.4 Conclusion 34
4. Legal Framework 36
4.1 Method 37
4.2 Laws & regulations 37
4.2.1 Participatiewet 37
4.2.2 SUWI 38
4.2.3 GDPR 39
4.3 Similar cases 41
4.3.1 SyRI 41
4.3.2 Sleepwet 43
4.3.3 GGZ against NMD municipalities 44
4.4 Interview lawyer of Gemeente Amersfoort 44
4.5 Legal framework 46
4.6 Conclusion 49
5. Algorithm Versus Human Experts 50
5.1 Method 51
5.2 Data understanding and preparation 52
5.3 Understanding the algorithm 55
5.3.1 Information split 56
5.3.2 Gini index 57
5.3.3 Random Forest 58
5.3.4 Lasso 60
5.3.5 Rulefit 61
5.3.5 MARSplines 62
5.3.6 Conclusion of the algorithm's results and reasoning 63
5.4 Understanding the human experts 68

5.4.1 Data preparation 68
5.4.2 Interviews 69
5.4.3 Conclusion of the human experts' results and reasoning 71
5.5 Conclusion 72
6. Ethical Framework 75
6.1 Method 76
6.2 Brainstorm[BR] 76
6.3 Project framework 77
6.3.1 Beneficence 79
6.3.2 Non-maleficence 80
6.3.3 Autonomy 83
6.3.4 Justice 83
6.3.5 Explicability 85
6.4 Validation 89
6.4.1 SyRI 90
6.4.2 Performance of the project framework on SyRI 92
6.4.3 Amersfoort 96
6.4.4 Performance of the project framework on Gemeente Amersfoort 99
6.5 Conclusion 102
7. Effectiveness, Accuracy, Legitimacy & Justness 104
7.1 Method 105
7.2 Effectiveness 105
7.3 Accuracy 107
7.4 Legitimacy 110
7.5 Justness 111
7.5.1 The algorithm should be free of bias and discrimination 111
7.5.2 The data that is used is proportional to the goal 112
7.5.3 The data should also be relatively privacy friendly for people with background knowledge or when combined with other data sources 113
7.5.4 The algorithm should make efficient use of resources 113
7.5.5 Conclusion 113
7.6 Implications for the design of fraud detection algorithms 114
7.7 Conclusion 116
8. Application to Gemeente Amersfoort 118
8.1 Method 119
8.2 What is the current situation regarding fraud detection at municipalities? 119
8.3 What criteria of fraud detection algorithms make them lawful to be used? 119
8.4 Which ethical considerations should be taken into account when using a fraud detection algorithm? 121
8.5 In what ways is algorithm-based fraud detection different from human expert-based fraud detection? 121

8.6 Are the conventional assessment criteria of the algorithm legitimate, justified, effective and accurate? What are the implications for the design of fraud detection algorithms? _____ 122

8.7 Conclusion _____ 122

9. Concluding Remarks _____ 123

9.1 Research questions _____ 124

9.2 Discussion and future research directions _____ 125

9.3 Scientific contribution _____ 127

9.4 Practical contribution _____ 128

References _____ 129

Appendix _____ 132

Appendix A _____ 133

Appendix B _____ 134

Appendix C _____ 136

Appendix D _____ 139

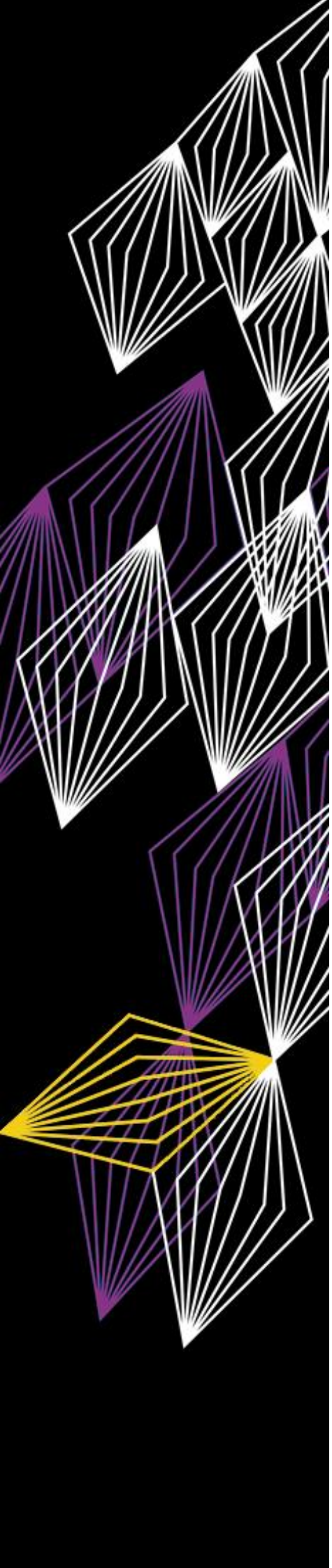
Beneficence _____ 139
Non-Maleficence _____ 139
Autonomy _____ 144
Justice _____ 144
Explicability _____ 146

Appendix E _____ 155

Beneficence _____ 155
Non-Maleficenc _____ 156
Autonomy _____ 161
Justice _____ 161
Explicability _____ 164

Appendix F _____ 175

Appendix G _____ 178



Chapter 1

1. Introduction

In this chapter, the topic is introduced, research questions are formulated and the structure of the rest of the report is given.

1.1 Introduction

More and more data are being generated in all aspects of human life. In 2018, 2.5 exabytes (= 2.500.000.000 gigabytes) of data were generated per day[1]. It is expected that this amount will grow to 463 exabytes in 2025 with the growth of the Internet of Things and the appearance of new services (Reinsel, Gantz, and Rydning 2018). Many opportunities arise with having an extensive amount of data, like customer profiling, predictions, and website optimisation. These data are so valuable that they are already considered a business asset and protected through trade secrets (Reddix-Smalls 2011). With the value and convenience of big data, also certain risks are involved, such as data leaks and the fact that large amounts of data may be difficult to discern (Zhang 2018). In 2012, the European Union recognised these (increasing) risks and created what would later become the General Data Protection Regulation (GDPR), which has been installed since May 2018[2].

Fifty percent of the municipalities in the Netherlands already use data to a certain extent and face the same opportunities and difficulties. 56% of governmental institutions perform research on algorithms (Doove and Otten 2018). The Rathenau Instituut[3] mentions some examples such as Utrecht, who uses data to manage their water pumps automatically and Eindhoven, who uses cameras to monitor their streets, predict (violent) behaviour that will occur there, and react to this. To use these amounts of data, some municipalities and other governmental institutions are creating fraud detection algorithms. These algorithms scan through individuals and create a subset of potential fraudsters. This subset can then be checked by a human who can then find out whether there are actual fraudsters in this relatively small subset. This is done because the damage of fraud is huge; estimated at more than 1 billion euro of benefits fraud in the Netherlands alone (Geldrop and Vries 2015) and because the available resources for human inspection are limited.

Privacy and ethical concerns arise, as seen by another fraud detection algorithm: Systeem Risico Indicatie (SyRI) from the Dutch government. This system received an ironic privacy prize for the invasion of privacy of people[4]. There were five reasons for this, as displayed in Figure 1.

In addition to this, parliamentary discussions were held about this topic recently[5]. There are thus indeed some public concerns about this system and therefore it is important that municipalities and other governmental institutions do as much as they can to minimize these concerns.

---

[1] https://www.domo.com/solution/data-never-sleeps-6

[2] https://gdpr-info.eu/

[3] https://www.rathenau.nl/en/digital-society/data-driven-cities

[4] https://www.binnenlandsbestuur.nl/sociaal/nieuws/syri-grootste-privacyschender-2019.11606010.lynkx

[5] https://debatgemist.tweedekamer.nl/debatten/vragenuur-257?start=622

Figure 1 - Reasons for the big brother award of SyRI

1.2 Problem statement

Municipalities in The Netherlands must prevent, detect and punish fraudsters according to the *participatiewet* ch. 1.2 art. 8b[6]:

> *"De gemeenteraad stelt in het kader van het financiële beheer bij verordening regels voor de bestrijding van het ten onrechte ontvangen van bijstand alsmede van misbruik en oneigenlijk gebruik van de wet en de daarop berustende bepalingen."*

Which roughly translates to 'Within the framework of financial management, the city council lays down rules for combating the unlawfully acquired social security assistance and the abuse and improper use of the law and the provisions based on this law.'

Municipalities often have only a small team of people to do this. Furthermore, not all types of fraud are equally detectable (and therefore punishable). Gemeente Amersfoort, for example, indicated that the possibility to be caught for undeclared work is 10%, while the possibility to be caught for cohabitation fraud is 50%. It should not be fair that someone has a slimmer chance of being caught simply because they chose the right type of fraud to

[6] https://wetten.overheid.nl/BWBR0015703/2020-01-01#Hoofdstuk1_Paragraaf1.2

perform. To fulfil this duty of finding fraudsters, municipalities should detect fraud in the reasonably most efficient way possible. Algorithms may create the pathway for doing so, but as briefly mentioned in the introduction, other concerns arise. Mainly the privacy side and ethical side are of concern.

The main objective of this thesis is therefore to create frameworks that municipalities can use to create and implement their fraud detection algorithm in a way that is both privacy and ethically friendly. Another objective is to analyse the use of a fraud-detection algorithm and thereby gathering more information about the process and gaining new insights.

1.3 Case objective

Gemeente Amersfoort is very interested in creating a social security fraud detection algorithm, as determined in a meeting held in January 2020. The primary reason that they want such an algorithm is that they want to uncover their blind spots. They feel that every type of fraud should be punishable according to the same percentages. Furthermore, they think they might be prejudiced by whom they chose to check upon, and they hope an algorithm will be more objective. Last, they would like to use the algorithm later on to stop the fraud as early as possible. This is important for detecting people who become socially isolated, because they want to help those people as early on as possible and they want to prevent the fraud from continuing, since this would have large negative effects on people. The latter is an additional goal, but not the main focus.

What they also mentioned was that they do not necessarily see difficulties in creating such an algorithm, as this is outsourced to Totta Data Lab[7]. They see most difficulties in communicating this algorithm. They are therefore interested in the ethical side of the algorithm and how they can explain this to people. As they said themselves, the algorithm is not very different from what they do now, only now they check people randomly or when they are suspects, but they 'often fish in the same pond over and over again'. An algorithm can increase this pond and only select 'suspicious fishes'. The final check is therefore not different, what is different is how they choose the subset of people to check. Explaining this, however, is rather difficult and they must be sure to have the support of some important stakeholders, such as the general public, the client council and the councillor. An important part of this is thus to make it very clear what the differences are between an algorithm and a human agent. Both in ways of working and in efficiency.

Furthermore, the project leader of Gemeente Amersfoort mentioned they must be sure to comply with the GDPR. They already did some things to ensure this, such as only using the social security dataset and not coupling this with other datasets, but a close look should be given to check whether this is sufficient. Amersfoort has their own lawyer, with whom was be cooperated to see all sides of this project.

The frameworks developed in this research were applied to the use case of Gemeente Amersfoort. They provided information and data to further improve this research and to

[77] https://www.Totta Data Labdatalab.nl/

validate this research. The case objective therefore is a specific analysis of the use case of Gemeente Amersfoort using the frameworks developed in this research, thereby detecting focus points in ethical and legal areas for their fraud-detection algorithm.

1.4 Research questions and method

The main question of this research is:

How can data-driven fraud risk assessment algorithms be designed such that they are legally privacy-proof and ethically justified?

To answer this question, six sub questions were created. The first question looked at how social security assistance fraudsters are detected without an algorithm and why municipalities should find fraudsters in the first place.

1. What is the current situation regarding fraud detection at municipalities?

This was done by interviewing important stakeholders at Gemeente Amersfoort. These include the law enforcers and the project leader. In the interviews was also looked at the differences between Gemeente Amersfoort and other municipalities.

Second, since it is important to know what data are used and where these data come from. If this has been found, the follow-up question is whether there is some foundation for algorithmic fraud detection and where this foundation comes from.

2. What criteria of fraud detection algorithms make them lawful to be used?

This was done by having interviews with important stakeholders, such as the information provision advisor and the lawyer of Gemeente Amersfoort. Furthermore, similar cases, such as SyRI, were reviewed to see what was done and whether that was legitimate.

Gemeente Amersfoort indicated they are interested to see what exactly are the differences between an algorithm and a human in terms of detection to gain new insights.

3. In what ways is algorithm-based fraud detection different from human expert-based fraud detection?

This was performed with a test dataset which can be tested with experts and with an algorithm. In this part there was looked at the differences in methods between the experts and the algorithm. This also gives some answer to why municipalities would choose to use such an algorithm.

Third, as a large part of the pre-research already showed and Gemeente Amersfoort indicated as well, is that the ethical side is a very important aspect.

4. Which ethical considerations should be taken into account when using a fraud detection algorithm?

For this question, a framework was designed that can be used by municipalities for their fraud-detection algorithm development and implementation. This is based on literature, on practical insights, and brainstorm sessions.

Fifth, the algorithm itself was looked at. Part of this is displayed in the ethical framework and should be considered more in depth in this question. Here, the question is whether the tags used for the algorithm are legitimate, justified, effective and accurate. Are more tags needed to get more accurate results, or can approximately the same results be achieved with less information per case (fewer attributes)?

5. Are the conventional assessment criteria of the algorithm legitimate, justified, effective and accurate?
    a. What are the implications for the design of fraud detection algorithms?

This question was tested with the anonymised dataset and an algorithm. Attribute selection was performed on the test dataset and the earlier developed legal and the ethical framework was re-evaluated with the newly gained information.

Last, the use case in this project is the one of Gemeente Amersfoort. Therefore, the last question was applied to this specific case.

6. What do the results of question 1-5 mean for Gemeente Amersfoort specifically?

This was answerable by first answering the other questions and then tailoring these questions to Gemeente Amersfoort. The legal and ethical framework were applied here.

This research was validated by expert interviews along the way. This was especially important for the ethical framework. Gemeente Amersfoort indicated their lawyer could be used as expert interviewee as well. Furthermore, the ethical framework was validated by applying it to two practical cases.

1.5 Design Method

As design science research method (DSRM), the method of Peffers et al. (2007) was used whose framework can be found in Figure 2.

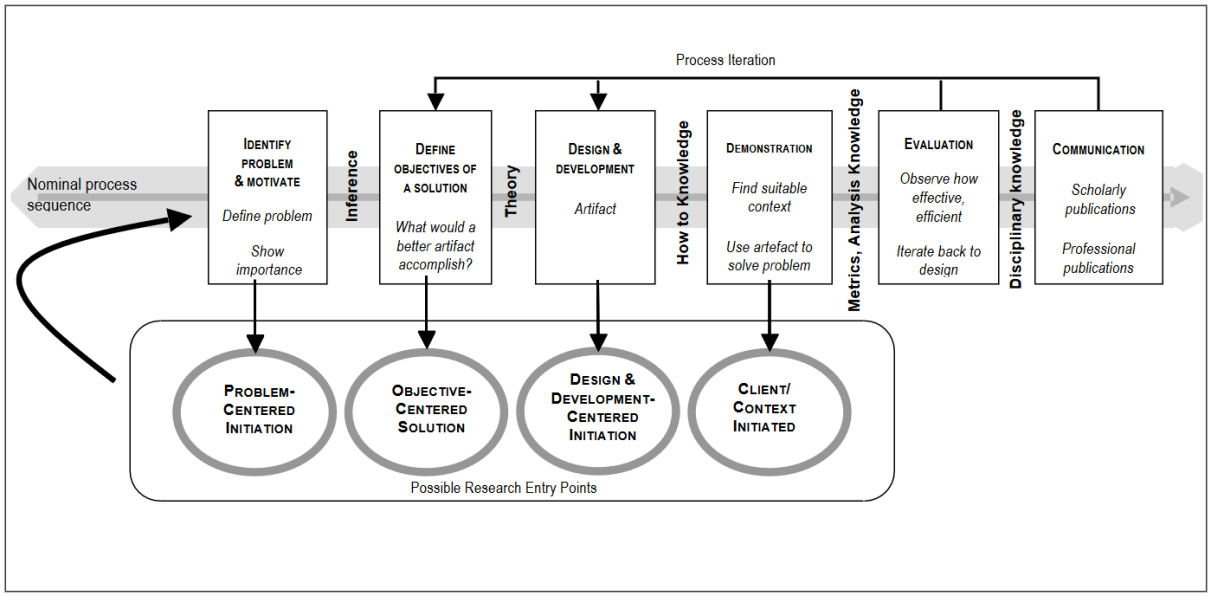


Figure 2 - DSRM of Peffers et al. (2007)

This method has been chosen because of its popularity and because it provides guidelines which are not too rigid, allowing for some freedom. Within this framework, iterations are still possible as well. The method is divided into six steps:

1. **Identify problem & Motivate**: The aim is to design a legal and ethical framework to improve the design and implementation of fraud detection algorithms in terms of privacy and ethics in the context of municipalities such as Gemeente Amersfoort. For this activity, background knowledge and the importance of its solution are needed. This is done in chapter 1, introduction, and 2, background information. Sub questions 1 and 3 also provide background information.
2. **Define objectives of a solution**: Sub questions 2 and 4 look at the background of the problem and at what should be part of the solution.
3. **Design & development**: In sub questions 2 and 4 a legal and ethical framework are developed, based on the gathered background information.
4. **Demonstration**: Demonstration of the ethical framework will happen in sub question 4 by applying it in two real world scenarios as validation. Sub question 5 focuses on further exploring the findings of sub-questions 2 and 4 by applying both frameworks to a real-world dataset from Gemeente Amersfoort, thereby expanding the insights on a data level. Sub questions 6 applies the legal framework to the real-world scenario.
5. **Evaluation**: The ethical framework developed in sub question 4 will be evaluated by applying it to 2 real-world cases. For sub question 2, the legal framework, an interview will be held with the lawyer of Gemeente Amersfoort. Both frameworks will be reiterated according to the findings from the evaluation. Evaluation takes place in each sub-question and in the conclusions chapter of this report.
6. **Communication**: This research will be communicated via this report which will also be available in Dutch. In the conclusion chapter final remarks are given as well as future research. Presentations will be held on the findings of this study and to expand on the implications of applying these frameworks in practice.

1.6 Practical note

While writing this thesis, the Corona crisis started. Apart from some general issues, this also had its impact on the structure and method of certain questions. How the questions were answered with the Corona crisis is already described above. In the discussion and future work, it is discussed what should have been done differently if the possibilities were there.

1.7 Structure

Chapter 2 provides the background information that is needed as a foundation before being able to answer the research questions and looks at related work. Chapter 3 answers research question 1 and provides extra information about the fraud detection process without an algorithm and expectations of stakeholders. Chapter 4 answers research question 2 by analysing related laws and regulations and similar court cases and validating this. Chapter 5 answers research question 4. In chapter 5, research question 4 is answered and the results are validated. Research question 6 is answered in chapter 7. Chapter 8 applies the results to the use case of Gemeente Amersfoort and, by doing so, answers research question 6. Finally, chapter 9 provides the conclusions found during this research, discussions and directions for future research, and it discusses the practical and scientific contributions.

Chapter 2

2. Background

In this chapter background knowledge is gathered to serve as a foundation for the rest of the report.

2.1 Municipalities in the Netherlands

2.1.1 General structure

To get a better understanding of the tasks of municipalities in the Netherlands, this subsection has been created. According to prodemos.nl, a website that explains all about Dutch democracy, municipalities have 11 main tasks[8]:

- **Civil affairs.** Knowing who lives in the municipality and giving out official documents (like a driving license).
- **Public order and safety.** The mayor has authority over the police and fire brigade, and each municipality has rules about cases like lighting fireworks.
- **Economics.** Accessibility of business parks is part of this, just like the opening hours of stores.
- **Social affairs and employment.** The municipality is responsible for getting as many people to work as possible or providing a monthly payment to people who are not able to work.
- **Care, welfare and public health.** The responsibility of care for the habitants of the municipality is also a responsibility of the municipality. Things like youth services (nl: *Jeugdzorg*) and municipal health services (nl: *GGD*) also fall within this task.
- **Asylum policy and integration.** Municipalities provide asylum seekers with a temporary residence permit or an indefinite permit after five years.
- **Education.** Municipalities make sure that schools have a building to be in and also that there is money available for scholars who need extra attention. They also check whether every child goes to school according to the Compulsory Education Act (nl: *Leerplichtwet*).
- **Spatial planning and public housing.** In destination plans, the municipality states how certain areas should look like. Next to that, they also monitor housing construction.
- **Traffic and transport.** Another responsibility is that roads are in an acceptable shape and they determine which roads are freely accessible for everyone and which for destination traffic only and so on.
- **Environmental Management.** The municipality is also responsible for the air quality and the collection of (separated) garbage.
- **Culture, sport and recreation.** Things falling within this task are for example having swimming pools, sports fields and nature reserves.

This, of course, does not mean that the municipality must *perform* all of this themselves. It mainly means that they are responsible for it. They can outsource these tasks to other parties.

These tasks are often also separated in certain domains according to the GEMMA architecture[9]. According to this, there are six domains:

- **Governance.** Focuses on administrative duties like management, strategy and accountability.
- **Social.** Focuses on the social side of society, this includes work, income, care, and youth.

[8] https://prodemos.nl/kennis-en-debat/publicaties/informatie-over-politiek/de-gemeente/wat-doet-de-gemeente/

[9] https://www.gemmaonline.nl/index.php/GEMMA_Domeinarchitecturen

- **Space.** Focuses on the physical environment, including the management of public space, and waste collection.
- **Public services.** Focuses on providing municipal products, often requested by residents or businesses (including travel documents and sports promotion).
- **Public order and safety.** Focuses on administrative enforcement under the responsibility of the safety domain chain and includes prevention campaigns and disaster exercises.
- **Support.** Focuses on supporting administrative and primary tasks (including housing, personnel management).

Gemeente Amersfoort has its organisation divided into 4 domains. These are the following, including the part of the GEMMA structure in brackets:

- **Social.** (social)
- **Physical.** Focuses on permits, city and development (space, public order and safety)
- **Service.** Focuses on taxes, customer contact and archive (public services)
- **Business operation.** Focuses on the legal side, IT and facility services (support, governance)

The main topic of this thesis lies within care, welfare and public health and the social domain.

2.1.2 Gemeente Amersfoort

Amersfoort is located in the centre of the Netherlands (see Figure 3) in the province of Utrecht. According to the bevolkingsregister, Amersfoort has 156.286 inhabitants (2019)[10]. This makes it the 15th largest city in the Netherlands.

Figure 3 - Location of Gemeente Amersfoort

2.2 Social security

Since this project lies within the context of social security assistance fraud, it is important to first know what is lawful social security assistance. In the Netherlands, people have a right to receive social security aid when they meet the following requirements[11]:

- Person is living in the Netherlands
- Person is above 18 years of age
- Person does not have enough income or personal capital to provide for their livelihood*.
- Person cannot receive another provision or benefit
- Person is not in jail

*Not enough income is according to a social grouping in Table 1.

[10] https://amersfoortincijfers.nl/jive

[11] https://www.rijksoverheid.nl/onderwerpen/bijstand/vraag-en-antwoord/wanneer-heb-ik-recht-op-bijstand

Table 1 - Capital limit social assistance 2020

Living situation	Maximum permitted capital
Shared household	€ 12.450
Single parent	€ 12.450
Single person	€ 6.225

There are around 400.000 people who receive social assistance (*specifically bijstand)* in the Netherlands (Ipsos 2018). Social security fraud can be described as receiving (more) social security assistance, while this should not be the case. Please note that receiving less social security assistance is not considered fraud.

People can also request special social assistance for extra costs that cannot be covered otherwise[11]. This money can be received as a loan or as a one-time gift. The costs must be sudden and urgent. Examples of this can be costs for administration (*nl: bewindvoering*), costs for legal counselling, but also costs for a new washing machine when the old one suddenly breaks down and not enough money is present.

2.3 (Fraud detection) Algorithms

To understand the problem, the definition of an algorithm must be clear first. An algorithm can be described as a recipe to solve a certain problem. What is important is that algorithms are not a new thing. They have existed for a long time already, such as Euclid's algorithm that was created 300 BC. Currently, many applications use algorithms, such as facial recognition, spam detection, product recommendations on sites like Bol and Amazon, and fraud detection. The definition of algorithm used by this research will be intelligent computer algorithms that in some way help with the decision-making process about individuals or groups of people, involving personal data.

The last couple of years, fraud detection algorithms are being used and created more extensively. This is often done because of the expected gain in efficiency and effectiveness, and because a human agent is more expensive than using Artificial Intelligence (AI). Another reason is that an algorithm can help humans make better decisions (Kimbrough, Wu, and Zhong 2002). This will be elaborated in section 2.6 . Often, these algorithms use a dataset which combines different data sources. For fraud detection, these can be, for example, data on healthcare, social assistance, and property. The datasets are then combined to find certain unusual patterns. For example, if person A receives social assistance, but does not cancel this when getting a new partner. Or when person B works 'under the radar' and doesn't get official salaries, but does receive financial aid (nl: *bijstand*). According to VanATotZekerheid, the Dutch covenant for insurance companies[12], there are four natures of fraud:

[12] https://www.vanatotzekerheid.nl/begrippen/fraude/

1. **Majoring:** Claiming more than actually happened. Someone broke the zipper of my backpack, but I claim the entire backpack is broken.
2. **Feigning:** Pretending to have something to get it insured. I say my backpack is broken, but it works perfectly fine.
3. **Staging:** Intentionally doing something to get the insurance. I break my backpack, because I do not like it anymore, so I will get a new one for free from my insurance contract.
4. **Lying/Withholding:** Not telling everything when starting the insurance. My backpack was already broken, but I did not tell so when making an insurance contract.

From this, both person A and B would fall within lying/withholding, as do most of the examples of social security fraud. Do note that social security fraud is not always intentional.

2.4 Artificial Intelligence

Algorithms can be very simple automated instructions, whereas AI, most of the time, contains sets of algorithms. Hence, AI cannot exist without algorithms, while algorithms can exist without AI. AI can also be used to do more than simply following automatic instruction, by, for example, also learning from its outputs and can therefore also come close(r) to human decision-making. That is also a reason why AI is becoming more interesting recently. In this research static algorithms and learning algorithms are distinguished. A static algorithms entails that the algorithm does not learn and is essentially just a program to check a couple of if-then or rule-based statements (i.e., 'If person A has most of their transactions done in Amsterdam, but claims to live in Rotterdam then this is suspicious'). A learning algorithm, on the other hand, learns from its output or from other factors and, hence, changes over time. This second type of algorithm tends to be more complex and less transparent, but they can discover hidden patterns that humans might miss, which may lead to a higher accuracy. These algorithms often also need more data, because they need to learn from new findings. When looking from an organisation perspective, learning algorithms are more difficult to implement, because it is harder to get all employees on-board, it is technologically more advanced (thus needs better resources in both manpower and technical power), and because these algorithms may change over time. The last point calls for good protocols and understanding.

According to Shubhendu and Vijay (2013) "Artificial intelligence is the study of ideas to bring into being machines that respond to stimulation consistent with traditional responses from humans, given the human capacity for contemplation, judgment and intention. Each such machine should engage in critical appraisal and selection of differing opinions within itself. Produced by human skill and labour, these machines should conduct themselves in agreement with life, spirit and sensitivity, though in reality, they are imitations". So, in short, that computers can do the similar (cognitive) tasks as humans, and sometimes even better. Machine learning goes further, by providing a system with the ability to learn without being specifically programmed to do so. Machine learning falls within the scope of artificial intelligence, as briefly explained by Ian Goodfellow, Yoshua Bengio & Aaron Courville in Holzinger et al. (2018) and can be seen in Figure 4. Later on, in 2.6, case-based algorithms refer to this. Deep learning goes one step further and are based on neural networks with complex architectures with multiple internal layers.

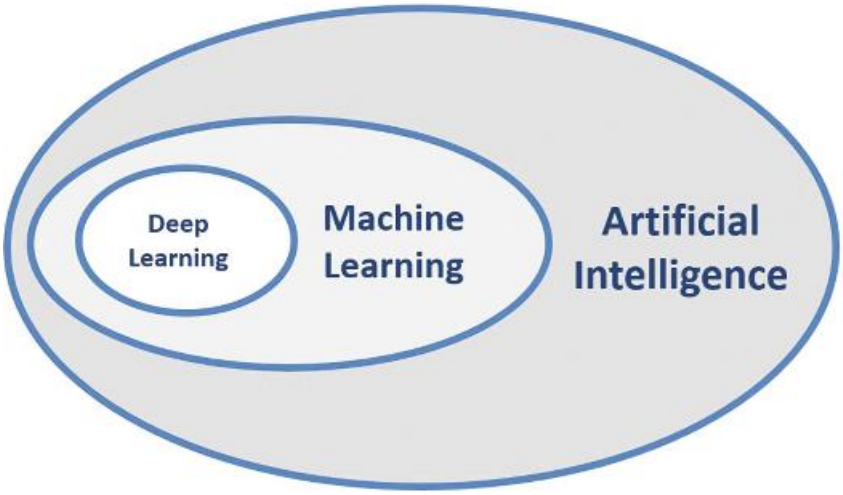


Figure 4 - Difference in AI, machine learning and deep learning according to Ian Goodfellow, Yoshua Bengio & Aaron Courville in Holzinger et al. (2018)

Advantages of AI lie, according to Shubhendu and Vijay (2013) in tireless performance of tasks, the ease of copying certain tasks, logical decision making (without emotions getting involved) and usefulness in hazardous environments. Disadvantages, again according to Shubhendu and Vijay (2013), lie in the risk of the AI breaking down, it losing data because of malfunctioning, the ethical and moral values, possible takeover of jobs and the fear of singularity. Another disadvantage lies in the possible bias that goes with AI[13]. This bias can both be learned from humans as from the learning process of the algorithm itself.

2.5 Stakeholders

It is important to know who are involved in the process of using algorithms within municipalities. Therefore, these stakeholders are identified for the use case of Gemeente Amersfoort. Involved with this project are 15 stakeholders. Who they are and what they do is listed below.

- **Project manager:** Is responsible for the planning and successful realisation of the project. They also are the direct contact person from Gemeente Amersfoort for this thesis.
- **Information provision advisor (*nl: IV adviseur)*:** Ensures the right data and applications are used. They look at the project from the data point of view.
- **Data scientist:** Looks into the algorithm of Gemeente Amersfoort.
- **Supplier:** This is Totta Data Lab in this case, providing the algorithm as a solution for fraud detection.
- **Legal advisor:** Works within Gemeente Amersfoort to make sure everything is legally allowed.
- **Communications advisor:** Give advice regarding the communication from the Alderman (*nl: Wethouder)*.
- **Enforcers**: Research potential fraudsters and track them down. There are six enforcers working with Gemeente Amersfoort.

[13] https://becominghuman.ai/amazons-sexist-ai-recruiting-tool-how-did-it-go-so-wrong-e3d14816d98e

- **Citizen managers:** They work at the municipality and process the applications for social assistance and check their justice. They are also the first point of contact for the customer (i.e., people who receive social security assistance).
- **Functional management:** Are the application managers; they have access to the dataset and provide the enforcers with information.
- **Team manager:** Manages the team, is also involved with this thesis.
- **Department manager**: Manages the department Job, Income and Care (*nl: Werk, Inkomen en Zorg)* in this case. The team manager falls within this department as well. They gave the order to start this project.
- **Alderman:** Is the public director within a municipality. They are comparable to ministers within the Cabinet, they are connected to a political party and are connected to a certain topic. They decide upon continuing this project or not.
- **Citizens of Gemeente Amersfoort:** Their data will be checked via the algorithm eventually, so they are involved with this project. People who do get social security assistance are called customers.
- **Client council:** In this council are customers of Gemeente Amersfoort. These can be organisations like the Federatie Nederlandse Vakbeweging[14] (FNV), social assistance customers and representatives.
- **Municipality council:** Consists of democratically chosen representatives of the municipality
- **Society:** They benefit when as many fraudsters are caught, because the tax system then works in the right way. They also gain trust in the government when fraud is stopped.
- **Opposing parties:** There are some opposing parties who actively try to stop these types of algorithms.

There will be meetings with the first four stakeholders at the beginning of this research. Meetings with the Information provision advisor, data scientist, enforcers, alderman and client council will be held later on.

2.6 Background of using algorithm within municipalities

Before starting, it is good to see what has already been done in this area. It must be kept in mind that much of the research by municipalities is kept as internal documentation and not open access. Besides, the related works are restricted to Dutch documents only, since municipalities might have a different role in other cultures or the privacy regulations may differ.

2.6.1 Why do municipalities use fraud detection algorithms?

Doove and Otten (2018) mention two reasons for using these algorithms. First, they say that most governmental institutions use these algorithms to find risk cases. Second, algorithms enable the use of limited capacity of municipalities to detect risk cases, so this capacity can be used as efficiently as possible. They do, however, mention that there might be negative consequences as well, but that most respondents do not state these. These negative consequences might lie in that an algorithm is not always transparent, might give false positives or false negatives, can lead to resistance of employees, can cause discrimination

[14] https://www.fnv.nl/

and can lead to an invasion of privacy. The possible invasion of privacy may also create suspicion amongst citizens (van der Weerdt and de Vries 2014). The Nederlandse Omroep Stichting (NOS) emphasises the, according to them, inherent discrimination and transparency issues, but also the efficiency gain[15]. Doove and Otten (2018) also mention that the choice to use an algorithm mostly depends on its explicability, testability and accuracy. Privacy is also mentioned as an important aspect. A report from TNO[16] (van der Weerdt and de Vries 2014) adds to this that it can be relevant for municipalities to profile their citizens to create order in chaos.

2.6.2 How far are municipalities with this?

From the meeting with Amersfoort on 31-01-2020 was found that the Centraal Bureau voor Statistiek (CBS) is working on a framework that organisations can follow when implementing a fraud detection algorithm[17]. This is, however, not finished yet. They published an explorative research in November 2018 (Doove and Otten 2018). Algorithms mentioned in this research were described as 'intelligent algorithms used for production or research and the basis of decisions about human beings or otherwise have an impact on an individual or group of people'. They explored how many and which organisations make use of algorithms and which do not (yet). From here we find that 50% of the municipalities they surveyed already use an algorithm. They also distinguish between rule-based algorithms and case-based algorithms. Rule-based algorithms make decisions based on certain rules and can be viewed as top-down. Case-based algorithms work the other way around and train themselves based on cases and are mostly decision trees, gradient boosting, regression models and deep convolution neural networks. 16% of all algorithm-using governmental organisations asked, use rule-based algorithms, 37% use case-based algorithms and 47% use both.

2.6.3 Case example: SyRI

SyRI (Systeem Risico Analyse) is an instrument used by the Dutch government to detect potential fraud cases in the social domain. It uses and couples many different data sources from different organisations to prevent and detect fraudsters. SyRI is notorious because of its media presence[18] (see also Figure 1 for the main concerns about SyRI according to the media) and shows similarities to other algorithms in the social domain. SyRI profiles citizens to predict potential fraud, misuse, and other offenses. There is a discussion between two camps, people who think algorithms will work better than humans, because of the efficiency and objectiveness[19], and people who think it is an invasion of privacy[18]. TNO did a quick scan

---

[15] https://nos.nl/artikel/2286848-overheid-gebruikt-op-grote-schaal-voorspellende-algoritmes-risico-op-discriminatie.html

[16] https://www.tno.nl/en/

[17] https://www.cbs.nl/nl-nl/corporate/2019/41/tno-cbs-werken-aan-transparant-en-toetsbaar-ai-gebruik

[18] https://bijvoorbaatverdacht.nl/wat-is-syri/

[19] https://www.netkwesties.nl/1367/systeem-syri-heeft-minder-vooroordelen.htm

on this system (van Veenstra et al. 2019) and from here was found that using the algorithm led to good financial results for the government. However, two challenges were mentioned, namely that innocent citizens may be checked because of false positives and that there is a consideration between invasion of privacy and the efficiency of the algorithm. In February 2020, a court case was held on SyRI. The entire court case will be analysed in chapter 4, but in short, SyRI was deemed to make too much impact on the privacy of individuals involved and was not sufficiently transparent and verifiable. Hence, it was judged SyRI could no longer be used by the government. All troubles regarding SyRI will be analysed in chapter 4 (legal) and 6 (ethical) of this report, this part serves merely as an introduction to SyRI.

2.6.4 Conclusion

Many governmental institutions seem to be working on using algorithms. However, not much of this research is public or complete. The reasons for implementing a fraud algorithm are to fulfil the municipality's duty of finding potential fraudsters, efficiency gains, and to create order in chaos. There are some things to keep in mind, such as privacy, transparency, efficiency and accuracy, possible discrimination, and to convince employees too. The roadmap provided by ECP might help in providing guidance for implementing AI. Last, SyRI shows some of the concerns in a real-world example.

2.7 Ethical frameworks

One of the larger concerns for Gemeente Amersfoort is the ethical aspect. That ethics are becoming more and more important can be seen because a major consultancy firm (KPMG) mentioned AI ethicist as an essential position for good AI implementation[20]. In this part, the existing frameworks and (ethically) related work will be analysed. Most of these frameworks are not specifically made for governmental institutions, but they are still of use for this research. In chapter 6, these will be combined to create a project framework.

2.7.1 AI4People

AI4People is a framework on ethical AI created by Floridi et al. (2018). In here, six documents on responsible AI are analysed to find the common ground. In the complete research they describe risks and opportunities as well as twenty concrete recommendations in assessment, development, incentivisation and support. Their framework comprises five ethical principles, as seen in Figure 5.

[20] https://revistaidees.cat/en/thinking-about-ethics-in-the-ethics-of-ai/

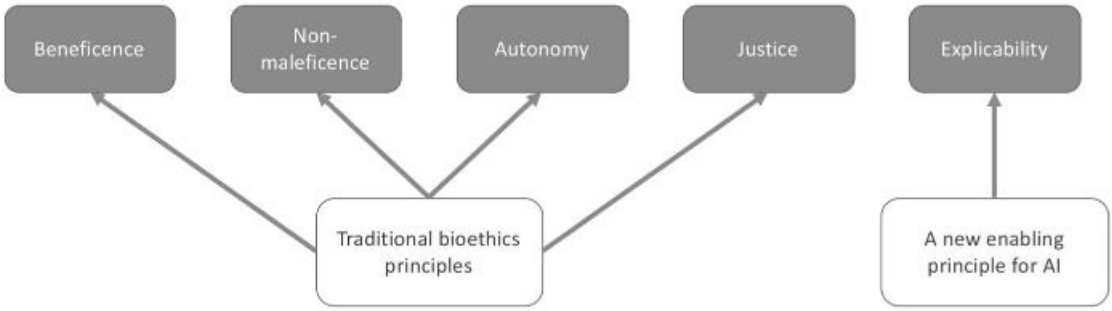


Figure 5 - Ethical framework AI4People

The first four principles were found from literature. First, beneficence is where humans' and planet's wellbeing are put upfront. Second, Non-maleficence is closely related to the beneficence principle. For this principle 'do no harm' is important, which entails not infringing the privacy of people and avoiding misuse of AI technologies in other ways. Most important for this principle is to prevent harm, whether from the intent of humans or the unpredicted behaviour of machines (Floridi et al. 2018). Autonomy is the third principle, where people have the right to decide. That this is not always the case is explained by a medical example, where a patient may not have the mental capacity to decide upon treatment and will be given the treatment, anyway. Here their autonomy is infringed, but it is not necessarily a bad thing. Besides this choice, there should also be a possibility to turn off the AI to regain control. The fourth so-called classic principle is justice, where no discrimination should occur. In a broader sense it is said that AI should respect the interests of all parties that may be affected by AI advances. They also added a fifth principle: explicability. This term is a combination of transparency and accountability and includes the questions 'how does it work?' and 'who is responsible for the way it works?'.

2.7.2 Guidelines for using data from online networking sites

Van Wynsberghe, Been and van Keulen (2013) created guidelines to provide a method for early assessment of ethical privacy concerns for researches dealing with social media data. The exact topic might not be perfectly aligned with this thesis, but it touches upon it. In this research, five guidelines are created, as displayed in Figure 6.

1. Make explicit the key actors: direct and indirect subjects, researchers etc.
2. What is the context and what does privacy mean in this context? (location and data content)
3. Type and method of data collection (passive vs active)
4. Intended use of info and amount of info collected
5. Value Analysis: making explicit and scrutinizing intended values of the researchers.

Figure 6 - Guidelines of van Wynsberghe, Been and van Keulen

The first guideline is to get a good overview of the stakeholders, such as the users, developers and system owners. Second, it is important to look at the context and privacy within this context. To stay within the social media domain, there is a difference between the privacy of a message posted in a private, closed Facebook group and a message posted openly, with hashtag on Twitter. Even if a message is posted in an open group, this does not mean that it can be used for anything. This is related to having a private conversation in a pub. If someone overhears this private information, it does not mean they can use it, or one step further, record it (as what social media analysis actually does). Third, type and method of data collection must be considered. Here it is looked at whether the data is collected with consent and whether it is collected actively (following people on a defined list created beforehand) or passively (follow everyone who comes by). The fourth guideline is the intended use of information and amount of information collected. Last is the guideline about value analysis. Here the researcher goes beyond the direct usage to see what the effect is. With fraud detection, as mentioned by van Wynsberghe, Been and van Keulen (2013), there is the positive side of having a fairer system, more financial security for the Dutch state and the fact that it is important to battle fraud and maintain fair taxes in this way. The negative side lies then in non-fraudsters still being checked by an AI, leading to a possible infringement of privacy. This is known as the value trade-off of justice and privacy, as also briefly mentioned in section 2.6. The intentions of the researchers must also be looked at. There is, for example, a difference between a societal important topic as detecting fraud and using social media data for marketing purposes.

2.7.3 Partnership on AI

Partnership on AI has a mission to benefit people and society with AI by conducting research, organising discussions, sharing insights, and responding to questions from the

general public[21]. They bridge people from different backgrounds, such as academics, researchers and companies. They have created eight tenets for its members. These are:

1. We will seek to ensure that AI technologies benefit and empower as many people as possible.
2. We will educate and listen to the public and actively engage stakeholders to seek their feedback on our focus, inform them of our work, and address their questions.
3. We are committed to open research and dialogue on the ethical, social, economic, and legal implications of AI.
4. We believe that AI research and development efforts need to be actively engaged with and accountable to a broad range of stakeholders.
5. We will engage with and have representation from stakeholders in the business community to help ensure that domain-specific concerns and opportunities are understood and addressed.
6. We will work to maximize the benefits and address the potential challenges of AI technologies.
7. We believe that it is important for the operation of AI systems to be understandable and interpretable by people, for purposes of explaining the technology.
8. We strive to create a culture of cooperation, trust, and openness among AI scientists and engineers to help us all better achieve these goals.

2.7.4 IEEE general principles

IEEE offers high-level general principles for creating and operating autonomous and intelligent systems (A/IS) (The IEEE Global Initiative on Ethics of Autonomous and Intelligent System 2019). These principles are listed below:

1. Human Rights–A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.
2. Well-being–A/IS creators shall adopt increased human well-being as a primary success criterion for development.
3. Data Agency–A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.
4. Effectiveness–A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.
5. Transparency–The basis of a particular A/IS decision should always be discoverable.
6. Accountability–A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.
7. Awareness of Misuse–A/IS creators shall guard against all potential misuses and risks of A/IS in operation.
8. Competence–A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

[21] https://www.partnershiponai.org

2.7.5 De Ethische Data Assistant (DEDA)

Dataschool Utrecht, in cooperation with data analysts from the city, created DEDA[22] to help data analysts, project managers and policy makers to recognise ethical issues in data projects, data management and data policies. The full framework can be found below, in Figure 7.

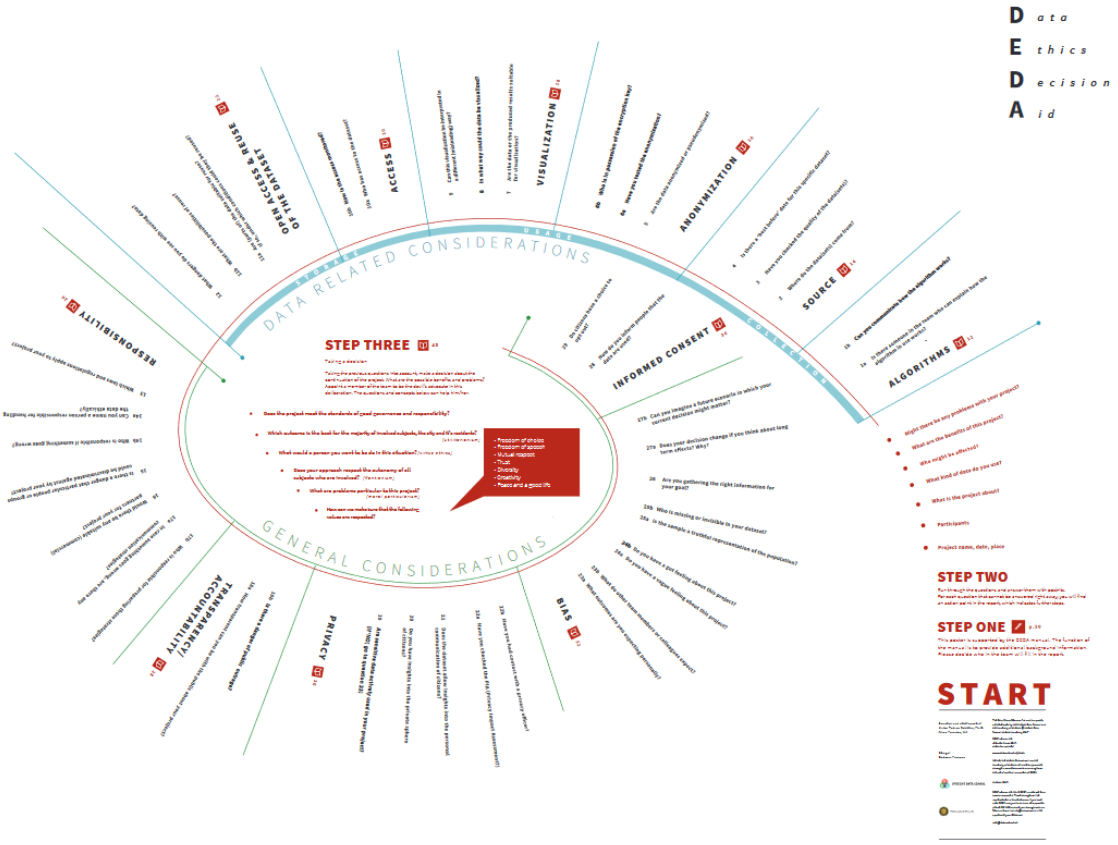


Figure 7 - Data Ethics Decision Aid (English version of DEDA)

In the first part of DEDA, the focus lies on the collection of the data. The algorithm itself should be clear and explainable and the data used should be of good quality . The second part focuses on the usage of the data. Attention is placed on anonymization of the data, and (if applicable) how to visualise the data. The third part questions about the storage of the data are raised. Who can access data and reusability of the data are points to keep in mind. After these three points, the framework goes on to more general considerations. First of them is the responsibility, which is about responsibility of protocols and communication, but also about prevention of discrimination. Second, transparency is mentioned, with inclusion of the question of how transparent one can be. Third, privacy is in the framework. This takes a more legal approach on whether privacy is considered, such as conducting a Privacy Impact Assessment and having contact with the organisation's privacy officer. The fourth part of DEDA mentions bias. Here the focus lies on whether the data sample is a truthful representation and on gut feeling. The fifth and last point focuses on informed consent, with

[22] https://dataschool.nl/deda/?lang=en

the question whether people can opt-out and how they are informed. This framework is more extensive than the others previously mentioned, because it also includes parts that are under the legal chapter in this research (for example, anonymization). The legal and the ethical part are closely related.

2.7.6 Artificial Intelligence Impact Assessment

Closest to a framework for fraud algorithms in municipalities is a relatively well-known framework for using AI made by the ECP | Platform voor de InformatieSamenleving (ECP 2018). This is rather similar to a Data Protection Impact Assessment (DPIA), which is sometimes obligated by organisations to perform. In here, eight steps to decide on whether AI would have a good use in an organisation are described. The full roadmap can be found in Figure 8. Most attention is given to ethical, privacy, and security considerations.

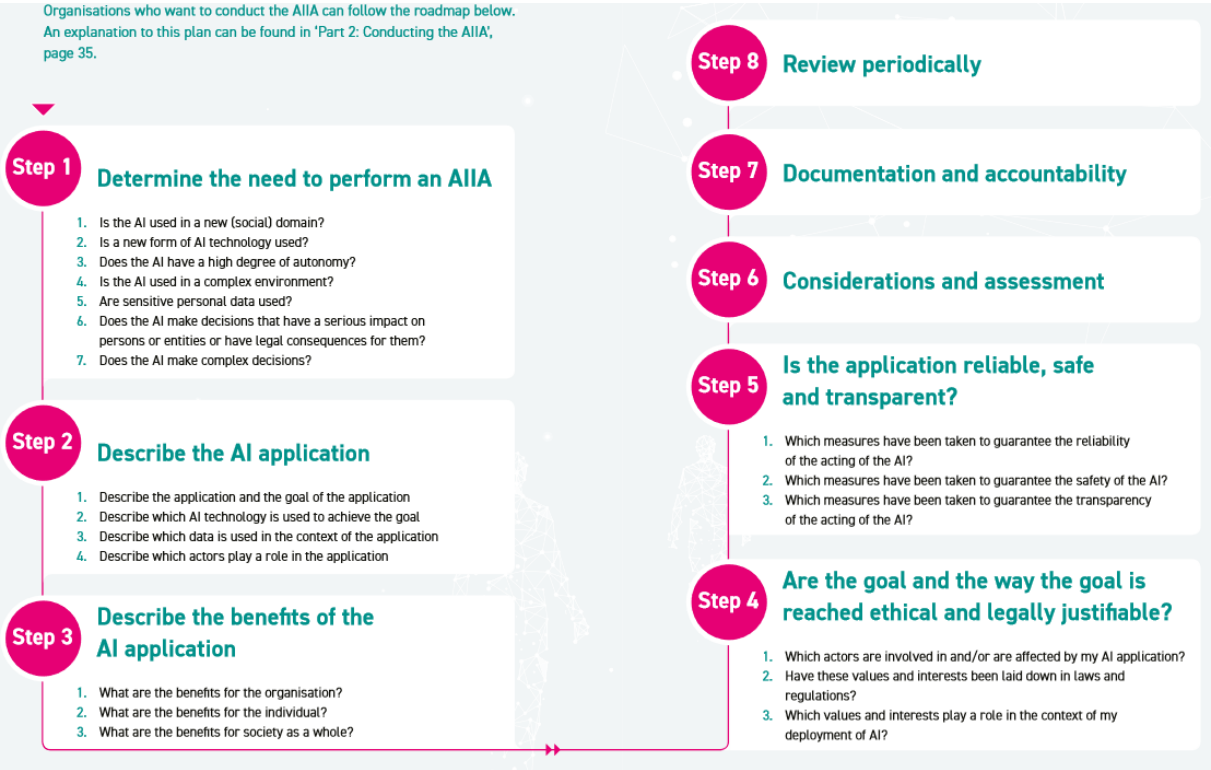


Figure 8 - Roadmap for Artificial Intelligence Impact Assessment (AIIA)

First, in step 1, it is questioned whether the AIIA should be performed at all. If this is the case, step 2 asks about general information about the AI application, such as its goals. If these are completed, in step 3 the benefits should be described, for the organisation, individuals and society. The 4th step is to check whether the goal is ethically and legally justified. Step 5 is to make sure that the application is reliable, safe and transparent. Then, step 6 is about the considerations that should be taken and the assessment. Step 7 ensures everything is documented and it is known who are responsible. Last, in step 8, it is ensured this impact assessment is reviewed periodically.

2.7.7 Other ethical frameworks

It should be noted that there are certainly more frameworks, such as the Asilomar AI principles[23] and The Ethics Guidelines for Trustworthy Artificial Intelligence[24], but they do not add new parts compared to the frameworks already mentioned. An interesting report is also published by the European Group on Ethics in Science and New Technologies, which takes a more philosophical view on moral dilemmas regarding AI (European Group on Ethics in Science and New Technologies 2018). However, because of the practical approach of this study, no elements from this framework were included, but it is still worth mentioning to any reader wants to go more in depth on that aspect.

2.8 Explainable AI

Gemeente Amersfoort indicated they are mainly looking at how to convince stakeholders about the potential of AI. This might be difficult, since AI is often represented as a black box (Goebel et al. 2018); data comes in, gets 'processed' and there is the result (see Figure 9).

Figure 9 - Black box programming [source Wikipedia.org]

What exactly happens in the black box is often not completely clear, both to the stakeholders as to the AI and its developers (Lipton 2018). However, it is an important step to make this clear in order to gain trust in AI, infer good causations and for transferability (Lipton 2018). Lipton also mentions that demands for fairness (as we have in the project framework) often call for interpretable models. These are generally agreed upon to be models that are understandable (transparent) or can be explained afterwards. Rai (2020) agrees that explainable AI can lead to more trust in the algorithm and fairness. He also mentions that the advances in explainable AI can make the trade-off of transparency and prediction accuracy less of a trade-off and more like a symbiosis. An even stronger stance is taken by Leenes (2016) who says that governments must step up their game, because they have something to explain to their citizens if they make decisions based on big data.

There are two main strands of work in explainable AI, according to Xu et al. (2019): transparency design and post-hoc explanation. Transparency design is about how a model functions, where the model structure, single components and training algorithms are understandable. As Lipton (2018) mentions, a person should be able to understand a model within a reasonable amount of time for a model to be called transparent. Interesting is that

[23] https://futureoflife.org/ai-principles/

[24] https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines#Top

Lipton warns that, in order to have transparent algorithms, we might need to sacrifice some quality of the algorithm. This is because some algorithms can surpass human thinking, greatly improving upon humans, but since humans cannot understand these, they can never be transparent. This does not seem very desirable, so there seems to be a trade-off.

Post-hoc explanation explains why a result is inferred, this results in analytic statements, visualisations and explanations given by examples. Lipton (2018) mentions here that we should not blindly accept these post-hoc notions, because these may be misleading (but plausible). For this, quite some techniques are already available, as explained by Rai (2020).

2.9 Fraud detection algorithm Totta Data Lab

To complete the background knowledge needed to understand the problem completely, an interview was held with Jesse Luk, co-founder of Totta Data Lab, to ask some questions about the algorithm of Gemeente Amersfoort. Benefits of using an algorithm that were mentioned by Totta Data Lab during this meeting were objectiveness, pro-activeness of detecting fraud, since it does not react on a signal, but searches actively in the data, and detecting more fraud than human agents would. Detecting more fraud results in retrieving more unfairly obtained social assistance funds and disturbing fewer people who do follow the rules. A drawback that was mentioned was that there might be bias present in the data, partly because the data that are used for the algorithm are real-world historical data (and law enforcers might not have been objective).

The difference between the algorithm of Totta Data Lab and, for example, SyRI lies in the purpose limitation. Amersfoort uses a subset of the dataset on social assistance, whereas SyRI used multiple data sources. Therefore, the algorithm of Totta Data Lab has the purpose limitation, according to them. The foundation of this algorithm is found in the GDPR and the participatiewet, whereas SyRI found its foundation in the so-called Wet structuur uitvoeringsorganisatie werk en inkomen (SUWI). Totta Data Lab was afraid that SyRI gave algorithms a bad reputation. This seems true, since there is a site (www.bijvoorbaatverdacht.nl) and there are many news articles against these algorithms, while there is not much in favour of these algorithms.

The algorithm of Totta Data Lab is written in R and Python, depending on what the local data scientist (of, for example, Gemeente Amersfoort) wants. The algorithm is case-based or learning-based using random forest or neural networks as a way of working. If there is a false negative and it is later found, through tips, that this was a fraudster while the algorithm did not detect this, the data will be fed back to the algorithm to learn from. The same applies for false positives. In this way it is hypothesised to eventually rule out as much of the bias as possible. Furthermore, it was stated that only lawful data are used that have purpose limitation. For example, addresses, nationality, birthplace or water usage are not included in the algorithm. Care should be taken with this, since certain combinations of data can still lead to indirectly using the same sensitive data. An example is given that data are used on a customer getting help with learning the language via a route by the municipality. Although these data are not directly data on ethnicity, it is highly likely that in this case their ethnicity is non-Dutch. What was intriguing is that it was mentioned that it would be interesting if

someone would file a lawsuit against them. They think they are more legitimate than SyRI and therefore would find it interesting to see what a court would rule about them.

It was mentioned that Totta Data Lab is working together with a scientist who focuses on explainable AI. Per prediction it is now possible to see which variables were most important. Next to that, they make sure their customers also receive the technical documentation of their algorithm. This seems to go more towards a combination of transparency design with the documentation and post-hoc explanation with the variables.

Last, it was mentioned that the algorithm for Gemeente Amersfoort is still being developed, however there are other municipalities that already use it. From these, three are public about using this algorithm, which are Gemeente Nissewaard, Gemeente Lekstroom and Gemeente Orionis. This might be interesting for a later stage.

Chapter 3

3. Fraud Detection Process

In this chapter, the process of fraud detection without the usage of an algorithm will be mapped. The results can be found in Figure 10.

3.1 Method

To discover the process of algorithm-free fraud detection (or the current situation), interviews will be held with one of the law enforcers, the project leader, the team manager and the alderman. This chapter also serves to discover more background information, such as percentages of fraud detection and possible benefits and drawbacks as seen by the interviewees. Finally, this chapter aims to answer research question 1: 'What is the current situation regarding fraud detection at municipalities?'.

3.2 Interviews

To get a better view of what currently happens when Gemeente Amersfoort tries to catch fraudsters (without an algorithm), interviews are conducted. Interviews are chosen as the research method, since this can create a good understanding of processes (Rowley 2012). Since there are only a few stakeholders on certain positions (for example, there are six law enforcers), questionnaires may not be too useful. Three such interviews are held to answer research question 1. First, the project manager is interviewed, to get an idea of the fraud-detection process. From this interview, questions can be added to the second interview. Second, one of the law enforcers is interviewed. They have to do the fraud detection process currently, so they know all the ins and outs of the process. Third, the team manager is interviewed to get a managerial view as well. A closed interview was conducted per mail with the alderman and is included in the text as well, since it serves to get a good insight into different views. This interview was conducted the beginning of April and for this interview it was not possible to ask further questions, since it was conducted per mail.

According to Rowley (2012), three types of interviews exist. First are structured interviews. These are rather similar to questionnaires; the difference lies in the timing in which a respondent must answer the questions. With questionnaires, they can answer at their own pace and hand the questionnaire in whenever desirable. With interviews, respondents must answer rather quickly. Questionnaires and structured interviews both require the possession of pre-knowledge to ask the right questions, because these questions cannot be altered after sending the questionnaire or creating the interview structure. Second are unstructured interviews. These types of interviews have a few focus points, but let the respondent talk relatively freely. Another aspect of these interviews is that the interviewer can adapt their questions to what was learned previously. Third, between the previous two types, are semi-structured interviews. These interviews contain six to 12 well-chosen questions in a certain order. Each question may have two to four sub questions which can be used by the interviewer to ensure that the respondent explores all desirable aspects of a certain question. The interviewee may also decide to ask questions in a different order (Knox and Burkard 2009) or go more in dept on certain questions if the interview goes into a certain direction (Hill et al. 2005). For these exploratory interviews to get knowledge about the process of detecting fraudster without the use of an algorithm, semi-structured interviews seem most fitting. This is because insufficient pre-knowledge has been acquired or can be acquired before these interviews, so structured interviews are not useful. Unstructured interviews also do not seem very fitting, since the necessary information still needs to be captured. Semi-structured interviews thus seem to be most appropriate.

Interviews will be conducted in Dutch, since all interviewees and the interviewer are Dutch. The interviews will be conducted in person. The interview setup can be found in Appendix A. The questions created were made with the guidelines of Rowley (2012) in mind. Hence, the questions should:

- not be leading
- not have two questions in one
- not result in yes/no answers
- not be too vague or general
- not be invasive
- be in self-evident order

The interview with the project manager can lead to a change in the designed questions and can be seen as a pilot interview. The results still will be considered. From this interview became clear what the focus points were for the other interviews. Namely, the focus for the law enforcer must lie in the process and the train of thought behind this. In the interview with the team manager, more focus was placed on the percentages and on the comparison with other municipalities. The interviews were held in the beginning of March 2020 and the results are described below.

3.2.1 Project manager

The project manager's function is to manage projects in which computerisation (*nl: informatisering)* plays a role. This is a rather broad task and can range from digitisation of forms to automation and process optimisation. The project manager is put into action by department managers to fulfil a project. This then starts with an exploration phase, creating a plan, and from this he makes sure it is executed correctly.

3.2.2 Law enforcer

In the interview with the law enforcer, it was asked what her job exactly entails. She mentioned that being a law enforcer differs from being an extraordinary investigating officer (*nl: BOA)*. She takes care of detecting social security fraud, with her main goal being to go from a story (signal) to real, objective facts to know what really happened.

3.2.3 Team manager

During the interview with the team manager, the focus was on the managerial side of the fraud detection process. His work task is to lead the six law enforcers and to have work meetings with them. Officially, there are seven law enforcers, the seventh is for the outer municipalities.

3.2.4 Alderman

The alderman is an important player in creating the algorithm, since in the end he must agree to it. Therefore, it is important to know his opinion and concerns. Within the municipality there are multiple aldermen who each are responsible for a different part. The alderman interviewed is responsible for Work & Income, Youth, Youth care, Diversity and Accessibility and Regional cooperation. Note that the interview with the alderman has been conducted via mail.

3.3 Results

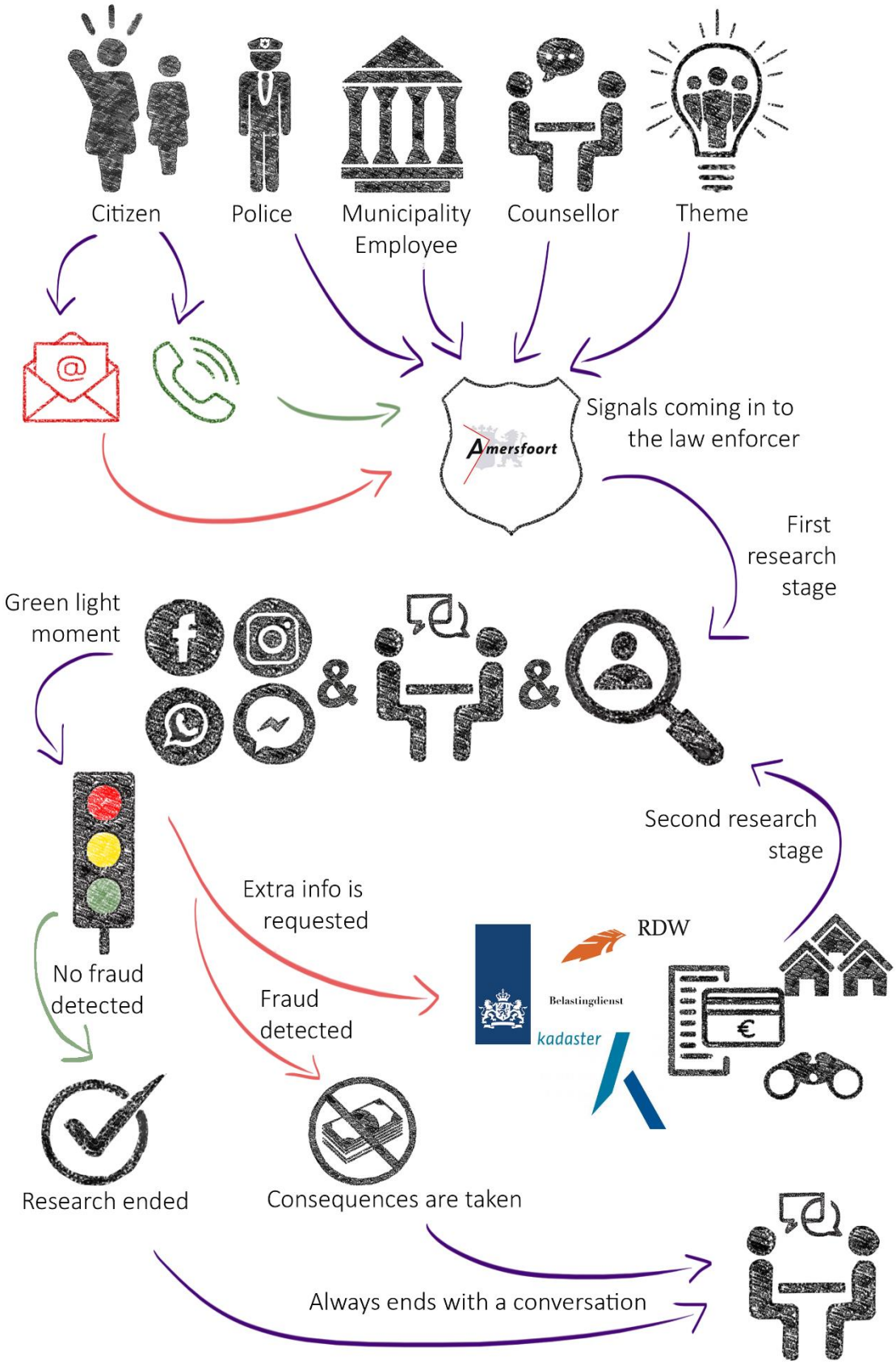


Figure 10 - Fraud detection at Gemeente Amersfoort

3.3.1 Fraud detection process

The full process of fraud detection can be found in Figure 10. Detecting social security fraud currently starts with a signal about a person (customer). This signal can come from 5 different origins:

1. **Citizens** can email or call, anonymously or not, to Gemeente Amersfoort when they suspect someone of being a fraudster. Tips per phone are preferred over mail because this gives the possibility to ask more questions. These tips are often from people in the near environment since they must know something about the person to give a tip about them.
2. **Customer managers & employees of the municipality** can also give tips. The customer manager's job is to process social benefit applications and to check their justification. When they see irregularities, they can alarm the correct persons within the municipality and thereby send a signal. Furthermore, all employees of the municipality can give a signal when they think something is off.
3. **The police** can give a signal when they suspect fraud.
4. **Path counsellor** (*nl: trajectbegeleider*) can give a signal. This, for example, happens when someone does not attend work for a few days in a row.
5. Via a **theme**, Gemeente Amersfoort looks at a certain topic. This is to focus on a group that might otherwise not be caught. An example of a theme is people who are receiving social assistance for over 3 years, but have never requested special assistance during this time. This is very remarkable, since social security is rather low and when something, for example, a washing machine or a fridge breaks, this money cannot cover this. These themes are created based on signals they often see during their own work (in-house) or in the media, or based on themes of other municipalities.

The fifth signal is the only one that is proactive, instead of reactive. If a signal is received, it is booked in the system. Then the law enforcer checks whether the signal can be transformed into facts. For this, they research for example social media, Kadaster (for property rights) and the Belastingsdienst (tax authorities) or they can talk with the customer's path counsellor. In this first stage, mainly the external sources that are already in the customer's file are used. It can be that these tips come from people who want to take revenge on the customer they are calling for, like ex-partners or fighting neighbours. At this first stage they do already inform the customer that they are being researched, even if they believe the tip to be unjust. After this first stage, for which the law enforcer has approximately 4 hours of research, there is a green light moment on whether to research the customer further. The case will be researched further if the law enforcer believes there may be a chance of fraud. This is an estimation. If this is the case, a meeting is planned with the customer and bank statements and extra information are requested. Here, the law enforcer can, for example, also talk with neighbours or do observations. What method is used depends on the signal that is coming in; for habitation fraud other things need to be researched than for illegal work. Also, there is started with the least impactful method and, depending on the severity of the (potential) fraud, this is scaled up to more impactful methods. For this second stage, the law enforcer has around 40 hours of research. From this research, a conclusion can be drawn, and this leads to a second meeting. From this second meeting, the conclusion often leads to the customer ending their social security assistance if needed. Gemeente Amersfoort does not impose a fine on the people who have received social security

assistance unfairly, as they do not think this is very useful since these fines often cannot be paid, anyway. They do impose a fine, however, when the fraud was intentional. From here, also assistance can be planned such as debt restructuring (*nl: schuldsanering*). Last, all this information is captured.

3.3.2 Background information on the fraud detection process

The most common types of fraud are undeclared work and habitation fraud (people who live together but do not mention this to the municipality) according to all interviewees. As explained by the law enforcer, for habitation fraud it is often the case that, for example, children are registered at another address, such as with family. This is so the parents get more money than when children are registered on their own address. Another type of habitation fraud is when people are subletting their house and live somewhere else in the meantime. The law enforcer also mentioned that Marktplaats fraud is a new, rising type of fraud. This is somewhat related to undeclared work, but it seems more of a grey area. With this type, people sell (and buy) items via Marktplaats (the Dutch eBay), but do not declare this. Luckily, municipalities can request data via Marktplaats to check this type, however at Gemeente Amersfoort they prefer to talk to the person first. The team manager added to this that owning a house outside of the Netherlands while not declaring this is also quite common.

When asked about the percentages, nearly every signal will get to the first research stage. Around 70% of the signals will be researched in the second stage. The percentages of the themes were not clear yet, since this is a work in progress. It was indicated by the law enforcer that tips from citizens less often lead to further research than tips from other sources. The team manager indicated he believes tips from citizen might be more useful, since they mainly entail real observations. However, he also mentioned that, since these tips are often anonymous, they cannot always be nuanced to get the best information. The team manager added that their goal is to get a 'fraud' result from around 40% of the cases that get past stage 1. Hence, they want 40% of people who were further investigated to result in fraud and the other 60% should be innocent (those are thus false positives for getting past stage 1). Interesting is that was mentioned that being a fraudster is not always intentional, some people did not know that they must declare some things (like using the car of your neighbour all the time). They use the conversations with customers also to inform them and keep these customers close. What was also noteworthy was that cases ending in a 'no fraud detected' verdict do not really mean that no fraud was committed, it may also mean that the fraud cannot be proven.

3.3.3 Preventing discrimination

The law enforcer indicated she does not necessarily want the percentage of innocent people to be checked (false positives) to be lower. She wants to keep the customer alert and maybe push them to get out of the social security assistance. She believes it to be important to let the customer know they are being thought of. The law enforcer also mentioned that she believed that someone may be checked, simply because they are receiving social assistance. The team manager agreed with this, because he does not want the law enforcers to only

take on cases, they know could be fraud. This may then lead to other cases not being detected, which is not desirable because this might lead to a biased pre-selection of potential fraud cases. Another reason he gave for not trying to get the percentage of false positives lower is that they want to make sure that the customer manager feels taken seriously when giving a signal.

The interviewees indicated that they actively tried to prevent discrimination on some fronts. They, of course, notice some cultural differences between cases. For example, some cultures are more monogamous than others (important for a theme on unknown parents they had previously). Law enforcers also try to check each other in conversations and address it to each other when they think someone is discriminating. Furthermore, they receive law enforcement trainings, not specifically on this topic, but it is included. Customers have the ability to mention that they feel discriminated. All conversation between law enforcers and customers are being recorded, so they can listen to these to check this. The latter does not check on discrimination in the process itself. It was indicated by all interviewees that they still try to keep it in mind not to discriminate, but there are no special ways to check this. Lastly, it was mentioned that there is no prototype fraudster, anyone can be one (intentional or not). There are however patterns that they see happening, which may be transformed into themes.

Something to keep in mind is that Gemeente Amersfoort is rather strict at the beginning of the process for requesting social security assistance (*nl: Streng aan de poort*). At this beginning, they already make an indication of the risk of this customer, for example, committing fraud or becoming socially isolated. These risks can be kept in mind when researching a customer. It also most likely makes the number of people committing fraud lower because of the high entrance criteria. Fraud can, however, also start after a while; therefore, it is still important to regularly check this.

3.3.4 Gemeente Amersfoort compared to other municipalities

When asked what the differences and similarities were between Gemeente Amersfoort and other municipalities, the law enforcer said that she believes that Gemeente Amersfoort is a precursor. She mentions other municipalities do less to find potential fraudsters or that they are very rigid. Gemeente Amersfoort tries to be a social municipality by talking with customers. For example, Gemeente Amersfoort seldomly visits customers only to check them. The team manager agreed with this; he also mentioned that since other municipalities can be very rigid, they spend more time on one case and can therefore check fewer people. This is not desirable for Gemeente Amersfoort, since they have the ambition to speak with every customer at least once a year. The law enforcer did however mention that this approach may not always be the best, since it may sometimes be too soft. The team manager further indicated that most other municipalities do not have the green light method. Most important for Gemeente Amersfoort in this process is to stop wrongfully acquired social assistance from continuing.

3.3.5 Benefits and drawbacks of a fraud detection algorithm

All interviewees were asked what they think about a social security fraud detection algorithm. The law enforcer was enthusiastic, because she believed an algorithm can make customers feel more seen and thought of. She also mentioned benefits for society. However, she was careful about taking old, previous cases into account. For example, she believed it is not fair that people who received social assistance ten years ago would still be marked by such an algorithm as potential fraudsters or risk cases (except for training). The project manager was also enthusiastic about this, since he believes AI is more efficient than humans and that AI may be more objective as well. He believes AI can help to overcome the differences between law enforcers (that one law enforcer might be stricter than another). The project manager sees drawbacks in the lack of transparency and that the process should be diligent. He mentioned an algorithm is 'High risk, High reward'. To the latter, the alderman agreed. If the algorithm works, it is very beneficial, since it shows that fraud doesn't pay. On the other hand, if mistakes are made, this can lead to public outrage. The team manager was also enthusiastic, because the algorithm does not make decisions, but only gives a risk indication. Like the project manager, he wants to minimise differences between people and thereby make it more fair for customers. He also wants more wrongfully acquired social assistance to be stopped as soon as possible. He finds it important to be able to know why someone is marked as a risk case, so explainable AI must be present. Last, he wants to extend the vision to find cases that would normally not be detected (i.e., 'increase the pond'). Interesting is that the interviewees all have different positions and different benefits and drawbacks identified partly based on this position.

The alderman has also been asked about possible benefits and drawbacks of an algorithm in a separate (mail)interview. As benefits he mentioned that fraud can be detected faster and easier, since fraud detection is rather complex and costs time. Furthermore, he mentioned that fraud detection can be done more effectively, since probably more fraud can be detected. He also mentioned drawbacks to the use of algorithms. Namely, the discrimination that might be present in an algorithm. He stated everyone should be treated equally, also by algorithms. Interesting is that he mentioned that discrimination might be present because these algorithms are created by humans. It was also mentioned that this might be solved by having a diverse team with different views. He highlighted the importance of privacy. He finds it important that all data that are used are stated clearly, such that it is clear which data are collected and why. Last, the importance of human interference was highlighted by the alderman. To do this, he really highlighted the importance of transparency and communication to and with the public.

Important to keep in mind with this part is that only one of the six law enforcers was interviewed. In terms of opinions, there are possibly some differences between them.

3.4 Conclusion

The fraud detection process of Gemeente Amersfoort has been uncovered and displayed in Figure 10. It is known that the process of Gemeente Amersfoort differs from other municipalities in that they are rather strict at the beginning of the process and fairly social

during the process. In addition, not all municipalities use the ‘green light’ system. However, the gist of the process will probably be the same. Benefits and drawbacks as seen by the interviewees were also found in this sub question. Benefits that were mentioned are that the customer might feel more seen, benefits to society because of detecting unlawful social security assistance, the efficiency, minimising differences between law enforcers (thus more fairness), detecting new fraud patterns and more fraud detection. Drawbacks (or points of extra attention) that were mentioned are the usage of old cases for training, lack of transparency, possible discrimination, invasion of privacy, there should be human interference, and the communication must be on point. If an algorithm would be implemented, it can be seen as adding another signal to Figure 10.

Chapter 4

4. Legal Framework

This chapter explores what criteria of a fraud detection algorithm make them lawful to be used to answer sub question 2. The most important points of these laws can be found in the conclusion or in the accompanied guidelines created from this in Figure 11.

4.1 Method

To create a legal framework, relevant laws and regulations will be analysed, as well as corresponding court cases. These will be combined into one framework. An interview with the law enforcer of Gemeente Amersfoort at the end of this chapter will evaluate this framework and complete it. This framework answers the question 'What criteria of fraud detection algorithms make them lawful to be used?'.

4.2 Laws & regulations

4.2.1 Participatiewet

To know why municipalities provide social security in the first place and what their rights and obligations are, the Participatiewet has been investigated. This was briefly done in the first chapters and will be extended here.

First, according to the Participatiewet §2.2 article 11.1, every citizen of the Netherlands who cannot provide for himself because of certain circumstances, has the right to receive social assistance.

> *"Iedere in Nederland woonachtige Nederlander die hier te lande in zodanige omstandigheden verkeert of dreigt te geraken dat hij niet over de middelen beschikt om in de noodzakelijke kosten van bestaan te voorzien, heeft recht op bijstand van overheidswege."*

There are exceptions to this, as listed in 2.2 Social security. Customers must inform the municipality about everything that might have an influence on their social security assistance, as mentioned in the Participatiewet section 2.3, article 17.1. Examples of this are declaring extra income and moving in with a partner. If municipalities have questions about this, the customer has the obligation to cooperate. Municipalities have the duty to prevent people from unfairly receiving social security assistance, according to Section 1.2, article 8b. Next to this, both the customer and the municipality try to get the customer reintegrated within society by offering for example extra trainings and by supporting customers in finding a job. Customers themselves also have the task to try to find work as soon as possible according to section 5.1, article 30a. Customers can also be asked to do social work for the municipality as a compensation (*nl: tegenprestatie*) and this will be obligated soon[25].

[25] https://nos.nl/artikel/2311340-tegenprestatie-in-de-bijstand-wordt-in-alle-gemeenten-verplicht.html

Furthermore, customers have the right to have input on the decisions of the municipality via the client counsel. This can be via themselves or via a spokesman who is in this counsel. This is described in the Participatiewet section 5.3, article 47. An interesting part is described in Article 53a, section 6 where is mentioned that 'het college' may check whether the data provided for receiving social security assistance is correct and complete.

Thus, the reason municipalities try to find fraudsters origins in the Participatiewet.

4.2.2 SUWI

In the SUWI decree, it is described how the different parties in the fraud prevention process can work together and exchange data to do so. Such a request for cooperation must be rather specified. These parties are, for example, the municipality, the Uitvoeringsinstituut Werknemersverzekeringen (UWV) and social work companies. Wet SUWI, Article 30a describes the reintegration process of customers, as described in the part about the 4.2.1 Participatiewet above too. Furthermore, in Article 31, it is described that for all customers it is indicated what their estimated employment prospects are. In Article 65, it is also described what information specifically the algorithm SyRI can use and combine. Examples of this can be data about a customer's job, water and energy usage and fines. The full list can be found in Appendix B. In Article 64 of the SUWI wet, it is described which organisations can make use of SyRI, of which municipalities are one. Later, in 4.3 Similar cases, a further look into the SyRI case will be taken.

What is good to add to this is that Article 64 and 65 were altered at the beginning of 2014 to capture the usage of SyRI in the law; this is called the SyRI law. Another thing to keep in mind is the definition of a risk notification according to Article 65 section 2, which can be roughly translated to:

> *"the provision of individualised information from the systeem risico indicatie [SyRI] containing a finding of an increased risk of unlawful use of government funds or government schemes in the area of social security and income-dependent schemes, taxes and social security fraud or non-compliance with labour laws by a natural person or legal person, and of which the risk analysis, consisting of coherently presented data from the systeem risico indicatie [SyRI], forms part."*

In Appendix II of the SUWI Decree, a list of information municipalities may use and share with other parties is specified.

4.2.3 GDPR

The GDPR is a relatively new regulation which came into place in May 2018 and displays the most important rules for handling personal data in the European Union. In the next part, important sections of the GDPR will be highlighted, such that a cohesive list of important points can be created for the conclusion of this chapter.

First, according to Article 5, personal data must be processed lawfully, fairly and in a transparent manner. The transparency principle requires accessible and understandable information, communication and plain language, and the provision of information to data subjects about the identity of the controller and the purposes of the processing. People should actively be provided information to make them aware of their rights and risks. Collection or processing of personal information must have a specific goal (purpose limitation) and it must only contain what is necessary (data minimisation). Whether the necessity is indeed there also depends on the costs of the system, the chance of being caught and the actual social need. If data are used that were not collected for the specific goal, a very close look should be taken on whether it is allowed according to Article 6, section 4. The data must be kept up to date, only be kept as long as needed and it must be secured properly. Gemeente Amersfoort in this case should be able to show that they comply with the aforementioned requirements. Article 5 mentions that customers must be informed when they are being researched. However, if Article 22 and Article 71 are taken into account, it is allowed to have monitoring to prevent social security fraud (Raat 2020). There must, however, be complied with Article 13, 14 and 15 in this case (see later paragraph).

According to article 6, the processing of these personal data is carried out in the public interest (namely, social security money goes where it should go and vice versa) and by an official authority (municipality) who must check on social security fraud. Even stronger is that municipalities must fulfil a lawful task, namely that of providing only lawful social security assistance and preventing misuse as stated in the Participatiewet and SUWI. This falls within Article 6, section 1c. The exact statement of this article is as follows: ‘Processing shall be lawful only if and to the extent that at least one of the following applies: (c) processing is necessary for compliance with a legal obligation to which the controller is subject’.

Article 13 highlights the importance of informing data subjects when data are being collected. In here, information such as about the processer, the timeframe during which the data is stored, why the data are being used and further rights of the customer should be mentioned (also seen in Article 15). If the data are not received from the customer, extra information on how the data was received must be added as well, as stated in Article 14.

As mentioned in Article 17, subjects have the right to be forgotten. In case of the municipality a timeframe of 5 years is handled, so fraudsters can be detected 5 years after they stopped received social security assistance. Until these 5 years, it is necessary for the municipality to keep the data. There are also fewer rights for the subject than, for example, with accepting cookies on a website. This is because the municipality must receive data in order to decide whether someone is eligible to receive social security assistance and in order to check them. This means that data subject cannot easily disagree with this and cannot ask to be forgotten before the 5 years are over. They also 'consent' by asking for social security assistance. Whether that is the ethical way of looking at it will be explored in chapter 6.

In Article 22, profiling is considered. Profiling, according to the general provisions means any form of automated processing of personal data comprising the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements. As stated in Article 22, decisions may not solely be based on automated processing, including profiling, unless the profiling is allowed by national law, then appropriate safeguards must be provided. These safeguards are difficult to fulfil when a learning algorithm or more complex algorithm is used, because of the (lack of) transparency of these algorithms and their complexity.

If there is worked together with another party who processes data, a processing agreement should be in place (Article 28). Article 30 stated that a register of processing activities must be created and kept up to date. Mentioned in Article 35, when processing is likely to result in high risk to the rights and freedoms of natural persons, a DPIA should be carried out. A DPIA should give a good overview of the risks of the data processing and the measures that should be taken by the party that processes the data. Such a DPIA should especially be done when automatic processing, including profiling is happening, when special data are used or personal data regarding criminal convictions and offences. These special data include racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health and data concerning a natural person's sex life or sexual orientation. In close relation with this, a data protection officer must be assigned as well, as stated in Article 38.

A good file for a rather complete overview of this topic is Toezicht op AI & Algoritmes by the Autoriteit Persoonsgegevens[26]. Furthermore, it is worth mentioning that there is being worked on a data usage and fraud detection (Wet gegevensverwerking door samenwerkingsverbanden), but as of now only a concept version is present[27]. This is taken into account when writing the conclusion, however.

[26] https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/toezicht_op_ai_en_algoritmes.pdf

[27] https://wetgevingskalender.overheid.nl/Regeling/WGK008727

4.3 Similar cases

4.3.1 SyRI

SyRI was already introduced in 2.6.3 in this report and will not be explained again, however this part will focus on the lawsuit against the law concerning SyRI. This was held on February 5th 2020 in The Hague[28] between a coalition of privacy focused organisations and the State. As briefly mentioned before, SyRI has its foundation in the earlier explained SUWI Decree.

SyRI has the following way of operating. When cooperation between parties has a good reason as a foundation and is approved for, the party assigned (in this case the Intelligence agency) combines the needed data and pseudonymises these data. Then the risk indication process sees which records (people) in the dataset are marked as high risk of committing fraud. In this process, points will be assigned to certain risk indicators and the higher the score, the higher the risk. These high-risk records are then decrypted such that the original person can be found again. Then in the second phase, the decrypted data is further analysed by the Ministry of Social Affairs and Employment whom decides if these records are worth investigating or not. If someone in the dataset is not marked as a potential fraudster, their data are deleted. If they are marked as a potential fraudster, the minister submits a risk report.

The court believes that new technologies must be utilised to find fraudsters, these may include self-learning algorithms. They state that finding fraudsters helps in citizen's trust towards the government, while the lack of transparency and lack of protection in general have the opposite effect. They are thus considering the dilemma of privacy versus the benefits of new technology. According to them, legislation must offer a sufficiently effective framework which allows the weighing of all interests in question in a transparent and verifiable manner. Privacy may be invaded if it is necessary for having a democratic society, proportionally again. The claim was that there is no proportionality with SyRI and that SyRI is not transparent. The defence of the State for SyRI not being transparent was that citizens could otherwise adjust their behaviour to not get picked out by the algorithm.

An important Article used during the lawsuit was Article 8 of the European Convention on Human Rights (ECHR). This article states:

> *"Everyone has the right to respect for his private and family life, his home and his correspondence."*

[28] https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBDHA:2020:1878

The second section of the same article mentions the following:

> *"There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others."*

The court thus had to judge on whether SyRI fulfilled the necessity part in this article.

The court concluded that the right of having a personal identity is closely related to the right of protection of personal data. Related to this as well is the right of respect for private life, in which protection for discrimination, stereotyping and stigmatising is included as well. Continuing on this, the court judged that the SyRI law impacts private life and therefore falls under Article 8 of the ECHR. When the SyRI law was created, this was also the conclusion, and the law was tested on whether it conflicted with, amongst others, Article 8 of the ECHR and the legislator judged it not to.

The SyRI legislation does thus infringe privacy, but the question arose in court if this is justified. Often mentioned by people very much against SyRI is that they would use 'any' data they can use, but the court judged that to be untrue, although the advice department mentions that there is almost no type of personal data to be thought of that does not fall within these categories. Data that can be used is specified are 17 categories, as can be found in Appendix B. The court mentioned that the money that gets 'lost' by fraud is very high (up to €1 billion) so there is necessity to do something, question then is whether SyRI is the good something. According to another case of *S. and Marper versus the United Kingdom*[29], as also discussed in the SyRI case, which has a more general part about data protection, protection must be offered against arbitrariness and the scope of the competent authorities must be clear. The court holds that the SyRI legislation in any case contains insufficient safeguards for the conclusion that it is necessary in a democratic society in light of the purposes of the legislation. A brief part of the court case was on whether or not SyRI made automatic decisions, and if so, whether this fell under one of the exceptions (of national safety or preventing illegal acts for example) of the ECHR article 8, section 2. The court did not judge on this.

So, SyRI combines multiple data sources to see whether someone might be a risk case. This does not necessarily mean that automatic decisions were made based on this algorithm and if so, whether this was justified. The risk model and indicators that make up the model and the data which are used in a particular SyRI project are not public, nor are they known to the data subjects, so there is no transparency as should be according to the GDPR. Furthermore, there is the fact that the data subject is unaware of the existence of a risk report, while the submission of a risk report has a significant effect on them. A customer is being informed at

[29] https://rm.coe.int/168067d216

the start of a SyRI project, but not when they are included in a risk report, which conflicts with the right on protection of personal data. The court is of the opinion that the SyRI legislation, insofar as it concerns the application of SyRI, does not strike the 'fair balance' required for the conclusion that there is a justified interference within the meaning of Article 8 paragraph 2 ECHR. Considering the principle of transparency, the principle of purpose limitation and the principle of data minimisation – fundamental principles of data protection – the court holds that the SyRI legislation is insufficiently transparent and verifiable to conclude that the interference with the right to respect for private life which the use of SyRI may entail is necessary, proportional and proportionate in relation to the aims the legislation pursues. The test whether data are necessary to be used was not performed by SyRI. Article 8, section 2 of the ECHR was not met, mainly because of the purpose limitation and data minimisation. The main articles that conflict with this are article 65 SUWI Act and chapter 5a of the SUWI Decree. These articles are judged to have no binding effect. This means that from this lawsuit on, multiple data sources may no longer be combined if there is no justifiable suspicion. A side note to this is that, because the court held to Article 8 of the ECHR, they refrained from answering some complex questions.

4.3.2 Sleepwet

Another case that was related to this topic were the summary proceedings about the Wet op de Informatie- en Veiligheidsdiensten (Wiv), also known as the 'Sleepwet'. This summary proceeding was held on 26 June 2018 between the plaintiffs, who are privacy concerned organisations such as Bits of Freedom and Free press unlimited, and the defender, the State. The act for Information and Security Services modernised the tools that they could use. In this law are, amongst others, the following parts included: Communication may be researched assignment wise, anyone may be asked to provide data, technical tools, false signals, false keys and third parties may be used to enter an automated work, and data may be shared with foreign services. The plaintiffs believed that the act, and mainly the parts highlighted above, led to an invasion of privacy. With this act, large amounts of data may be collected from innocent citizens. An invasion of privacy will be tested against Article 8 of the ECHR, as mentioned in the SyRI case as well.

Parts of this case are closely related to the SyRI case. There was for example discussed whether it was acceptable to gather data from innocent citizens. It was mentioned that it is unclear beforehand who is a threat and who is not. This is the same as with the social security fraud; it is not known beforehand who is a fraudster and who is not (with unintentional fraudsters in-between). Even though the research is not specifically targeting one person, it is not unfocused, but it is focused on one research topic. It should also be considered that further measures are taken, such as the security, retention period and proportionality. Hence, this was judged to be legitimate because it was 1) as goal oriented as possible, 2) during exercise of it, several essential principals of due diligence should be considered 3) special measures are taken to prevent these measures to be taken on confidential conversations between journalists and lawyers and the usage of data on confidential communication between lawyers and clients 4) the procedure is transparent, including the retention period and the deletion period 5) the system was tested on legality

by the Toetsingscommissie Inzet Bevoegdheden (TIB). This means that the legislator must perform many assessments before and after creating a new law. In short, it was decided that this law complied with the requirements of proportionality, necessity and subsidiarity. This means that there is a reason to check innocent people to find fraudsters as long as the proportionality, necessity and subsidiarity is present. Furthermore, the (entire) procedures must be described in a transparent way including retention and deletion periods and must be tested well by an independent party.

4.3.3 GGZ against NMD municipalities

On the 16th of December 2019, a case was held between the plaintiffs GGZ, Icare and De Trans, and the NMD municipalities[30]. The NMD municipalities and the GGZ must together offer care and the legal basis of the Social Support Act (*nl: Wet maatschappelijke ondersteuning, or WMO*) and the Youth Act (*nl: Jeugdwet*). The plaintiffs stated that the sharing of a list of personnel including educational data with the municipality of the person who should receive care was against the GDPR and the Procurement Law (*nl: aanbestedingswet*). In this part the focus will lie on the GDPR judgment.

The key similarity between this court case and this thesis is that it is also using personal data to execute lawful duties. It is not exactly stated somewhere that it should be done in this particular way (of sharing a personnel list), but this is a derivation of the law. It was ruled that this does not conflict with the GDPR. This was because the right of protection of personal atmosphere of life is not an absolute right, but should be seen in relation to other fundamental rights and freedoms. According to the court, they did fulfil the data minimisation requirement of the GDPR, and the purpose limitation. It was also judged that the necessity was present in this case. Interesting is that it was discussed whether the necessity was about the goal itself or about the processing of personal data for reaching the goal. The goals were compatible, namely the deployment of qualified personnel. The side note was made that special personal data should not be included here (and it was not).

4.4 Interview lawyer of Gemeente Amersfoort

On the 24th of March 2020, a semi-structured interview with the lawyer of Gemeente Amersfoort was conducted to validate the framework. Her job regarding this project is to dive into the case regularly to keep everyone wary of the legal (and ethical) side. She gives advice regarding the topic of data processing. She mentioned that currently a DPIA is being performed, and that she believes this is a good thing. This is mainly because not all the definitions and goals were clear and with this DPIA they should become clear. An example of a goal that was not yet clear is what the goal of using a fraud detection algorithm is. At first it was mentioned that it was to detect fraud, but the lawyer mentioned that the actual goal is to only provide people with social security assistance if they deserve to receive it and to stop wrongful social security assistance (in a way that is efficient) or to create better risk indications. An even higher goal is to get everyone back to work instead of relying on social

[30] https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:RBNNE:2019:5195

security assistance. Another example she mentioned was that it is unclear what fraud exactly entails. Is it fraud when someone sells 3 bikes via Marktplaats? And 15 bikes? Maybe it was not on purpose. These definitions and goals should therefore be clear before starting using the algorithm, and she hopes this will become clear by the DPIA.

What she also mentioned was that she sees benefits in using a fraud detection algorithm, because there is always subjectivity between different employees. If certain rules are made, every employee executes these rules differently. She also said that the dossiers are not always as detailed. Often it is only mentioned that the social security assistance stopped, but not necessarily why it did. This might be fraud, but it might also very well be that someone is back to work again.

One of the biggest downsides of the algorithm in legal terms as mentioned by the lawyer is the purpose limitation. When someone requests social security assistance, they have to provide lots of information to the municipality. The provide this information to get the assistance. The municipality receive this information to check whether someone is eligible to receive the assistance and also whether someone does not misuse their social security assistance, as mentioned in 4.2.3. Both parties provide and request the data not to train an algorithm. The algorithm, however, must be trained in order to stay effective. There is a discrepancy between these goals. Related to this is that old data, from customers who do no longer receive social security assistance should not be used at all for training the algorithm, according to the lawyer. The municipality may store social security assistance data from previous customers up to 5 years, to reclaim assistance if that proves to be needed. These data may not be used as training data, because the goal may differ. She mentioned that part of a learning algorithm can be compared to an experienced law enforcer, who learned from previous cases, but there is a difference between storing this information in someone's head versus on a computer.

Furthermore, it was mentioned that municipalities must check on fraud because of the Participatiewet. She added it is good to fish in the pond with the (suspicious) fishes. Very important according to the lawyer is the transparency of the algorithm. This transparency should be in such a way that people cannot outplay the algorithm, so there is transparency until a certain point.

In short, the most important points were that the algorithm should be transparent in such a way that it cannot be outplayed. The purpose limitation is very important; people do not provide data to train an algorithm, but to receive social security assistance. Historical data should not be included in the algorithm, since these people have no longer something to do with the municipality. Conducting a DPIA may help in creating clear definitions and goals. Compared to the earlier findings, this does add or emphasise two points of the framework: the DPIA and the purpose limitation.

4.5 Legal framework

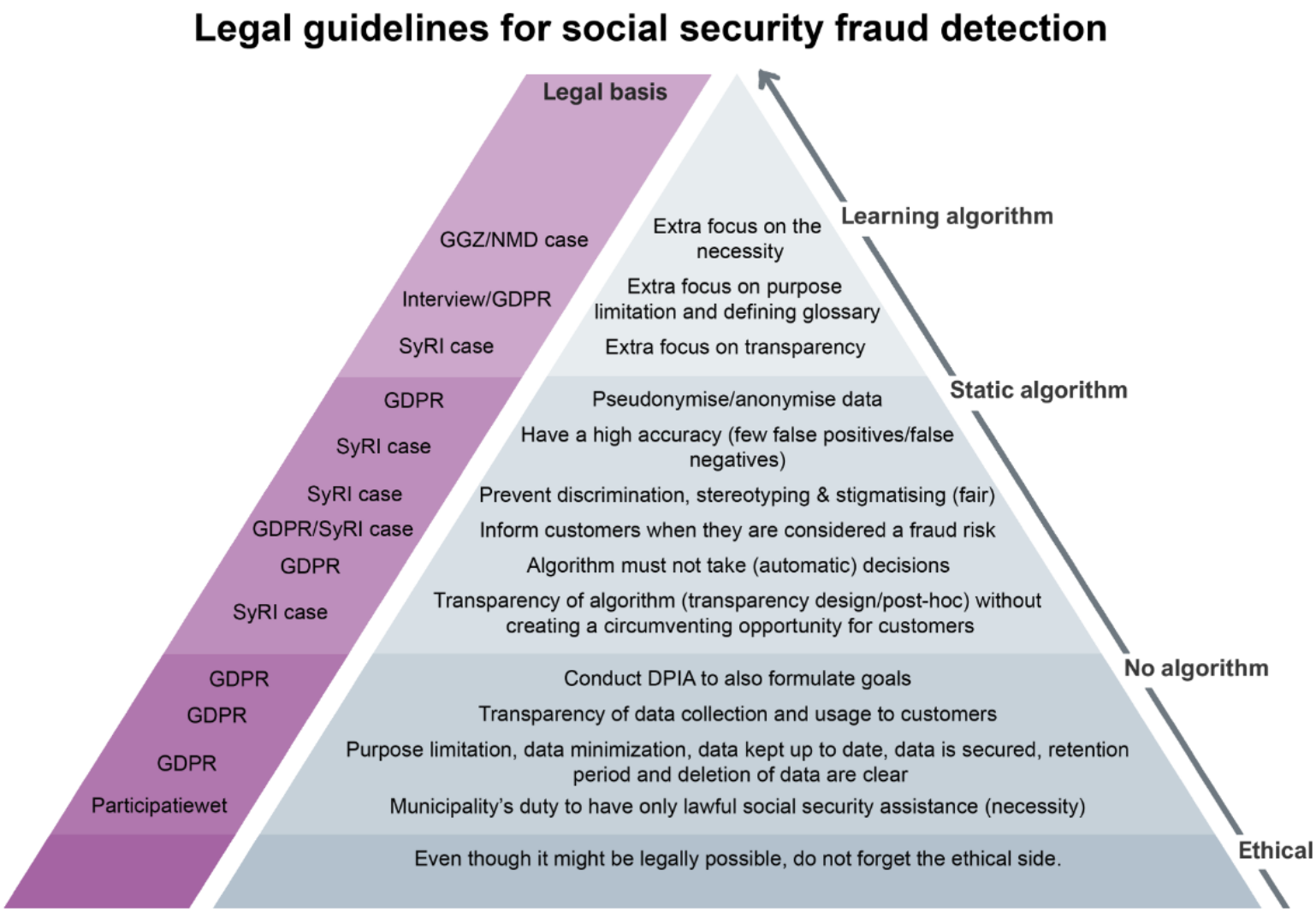


Figure 11 – Legal guidelines for social security fraud detection

From this part, a conclusion can be drawn on how to handle the legal part. This conclusion will serve as a set of guidelines, but will by no means replace the advice of a legal professional. The most important sections will be highlighted below, this does not mean that other parts of the law do not apply. The conclusion is summarised in Figure 11.

The municipalities indeed have the duty to prevent and detect fraud. This has a large societal importance. The lawfulness of this is lied out mainly in the Participatiewet. In SUWI it is specified which categories of data may be collected and used for this. From the SyRI case, it was found that detecting and preventing fraud is an important interest because of the large amount of money that goes with this and because it is important to have trust of the citizens in the government. The goal should be specified and documented by the organisation in such a way that the data can be used for an algorithm.

In terms of personal data, there are a few points that must be handled the same as they should be when no algorithms are used according to the GDPR. Data that are requested must be requested for a reason (purpose limitation) and there should be as little data requested as really needed (data minimisation). It should thus be necessary to give and have certain data. What data these are should be transparent for data subjects (for example, by informing them via a site or a flyer) and written up in a processing register. This also means that not all data sources can be coupled. An example of this is that the municipality needs financial data from a potential customer to check whether they are eligible for social security assistance. They, however, do not need to know on what they spent the 5 euros on 29-01-2019. A DPIA must be performed when the processing of data contains a high risk for the data subject. A DPIA is also good to formulate goals and definitions properly.

If the processing of data is performed by another party, a processing agreement must be in place. Furthermore, the data should be kept up to date and must be secured properly. These can be achieved by, for example, reminding customers of updating their data and ensuring all data are encrypted and keeping the data local. The data cannot be held forever, according to the lawyer of Gemeente Amersfoort social security data can be held up to five years after the contract ended (the customer stopped receiving social security assistance). After this period, the data must be deleted. These periods must be clear to the customers as well, which can, for example, be achieved via a site or flyers.

When using an (static) algorithm to discover fraudsters, some extra points come to the attention. What was the main issue in the SyRI case was the lack of transparency. It was not clear which data were used and why someone was a suspect. This was not allowed, which means that the algorithm should be as transparent as possible. This can be done via post-hoc explanation or by transparency design, of which both are feasible. An exception to full transparency was mentioned during the SyRI case, which stated that it should not create the opportunity for people to circumvent fraud detection because they know exactly on which criteria they will be checked. In any case, the entire process of fraud detection via an algorithm should not be kept a secret. Citizens should be informed when they are classified as being a fraud risk, which can be achieved by inviting them for a conversation. There is no necessity to inform customers when they are being checked in the first place, only when they are considered a fraud risk.

When an algorithm is used to check on fraud risk cases, discrimination, stereotyping and stigmatising must be prevented as much as possible. A human will also unconsciously do all these things, but the algorithm should be as objective as possible. Handles to achieve this could be for example by regularly evaluating the (fairness of the) algorithm. There should also be as few false positives and false negatives (accuracy) in the algorithm which can be done by giving the algorithm feedback, for example. The impact of being a suspect of fraud might be very high, since the municipality may visit their house or speak with their neighbour. This is also the case when fraud is checked without an algorithm, but it is still good to keep in mind. Next to this, data should also be pseudonymised or preferably anonymised when going through an algorithm. Lastly, no fully automated decisions may be taken based on (sensitive) data, especially when the algorithm is more complex. This can be done by simply letting a human interfere in the decision-making process.

When learning algorithms are used, extra focus should be placed on three parts. First on the transparency, since these algorithms tend to be less transparent by default (see 2.4 for a brief explanation). This can be done via post-hoc explanation or by transparency design, of which the first one seems most feasible.

Second, more focus must be placed on the purpose limitation and definition of the glossary. A learning algorithm must be trained. Often this is done by using real-world examples, such as old fraud cases. There is a problem with this, namely the purpose limitation. Customers provide the municipality with data to receive social security assistance and to be checked for this. They do not provide the municipality with data to train an algorithm. The GDPR mentions that reusing data is possible if the goals are compatible. Therefore, municipalities must put good thought into how to state their goal of using an algorithm. Suppose the goal for the algorithm of 'stopping fraud more efficiently' is chosen. A simple algorithm that is slightly better than humans already meets that goal, and a learning algorithm is not justified, even though it could have much better results. However, if the goal of using an algorithm is having as much legitimate social security, then goals are compatible with using an algorithm. The concept version of the Wet gegevensverwerking door samenwerkingsverbanden seems to agree on this[27]. Furthermore, customers who no longer receive social security assistance have the right to be forgotten after 5 years[31], but before these 5 years they could still be used for the algorithm if the goal is similar and formulated correctly. Careful thoughts should thus be put into this in collaboration with the organisation's legal department to decide on their viewpoint for their specific situation. Hence, the defining of the goal is a point that should not be passed too quickly. Part of this can be avoided by testing and developing the algorithm with fictional data instead[32].

Third, extra emphasis should be placed on why an algorithm requires certain data to be processed to achieve the stated goal(s). It is not completely sure whether this is the case for this topic specifically, since no explicit court case was held about this topic. There is also no

[31] https://vng.nl/sites/default/files/2020-02/selectielijst_20200214.pdf

[32] https://www.vka.nl/wp-content/uploads/2017/01/Whitepaper-privacyvriendelijk-testen-VKA.pdf

certainty whether necessity must specifically be laid out in a law or that it is free to be interpreted by the office. The closest case that was found on this topic (between GGZ and NMD municipalities) did rule in favour of sharing data to fulfil the (derived) legal goal/ This was a preliminary injunction, so another judgement may be made in the future. However, this gives an indication. Mainly because of these difficult issues with purpose limitation, it is not too clear cut whether algorithms can be used as a tool for detecting fraud. It depends on the interpretation of the law. What is a small comfort here is that the court ruled that new technologies such as (learning) algorithms should be utilised, but no exact judgement has been given on this topic specifically. Until then, the legality of using (learning) algorithms relies on interpretation of the laws and regulations. It is good to mention that even though it might be possible to make it work legally, the ethical side of it must definitely be taken into account as well. More emphasis on this will be placed in Chapter 6. This conclusion is summarized in Figure 11. The results of the validation with the lawyer of Gemeente Amersfoort are already altered in this displayed version. The reader should keep an eye on the updates for the new Wet gegevensverwerking door samenwerkingsverbanden.

4.6 Conclusion

In this chapter, a legal framework with guidelines has been created based on laws and regulations and similar cases. The GDPR, EHCR, Participatiewet, and SUWI have been reviewed. Furthermore, the SyRI case, Sleepwet, and the case of GGZ against the NMD municipalities have been analysed. The guidelines are categorised in 'no algorithm', 'static algorithm', and 'learning algorithm'. The framework is in pyramid shape, meaning one must work their way up. For example, if a learning algorithm is used, there must be complied with all underneath it. Meaning also the guidelines for no algorithm and static algorithm must be complied with. Important is that this framework especially highlights the importance of reviewing ethics in any case of data processing. These guidelines have been validated through an interview with the lawyer of Gemeente Amersfoort. The guidelines do not aim to report on all legal issues, it aims to highlight the most important issues, nor do they aim to replace the laws and regulations. If this framework were to be further used, more validation should be performed.

Chapter 5

5. Algorithm Versus Human Experts

This chapter explores the differences between humans and AI in fraud detection. The result of this chapter is summarised in Figure 20.

5.1 Method

In this chapter will be discovered what the benefits and drawbacks of a fraud detection algorithm are and how this would compare to a human approach. The exact content of a fraud detection algorithm and dataset may need to be kept a secret for obvious reasons. Aggregated results and other non-identifying results have been discussed with Gemeente Amersfoort, before they ended up in this report.

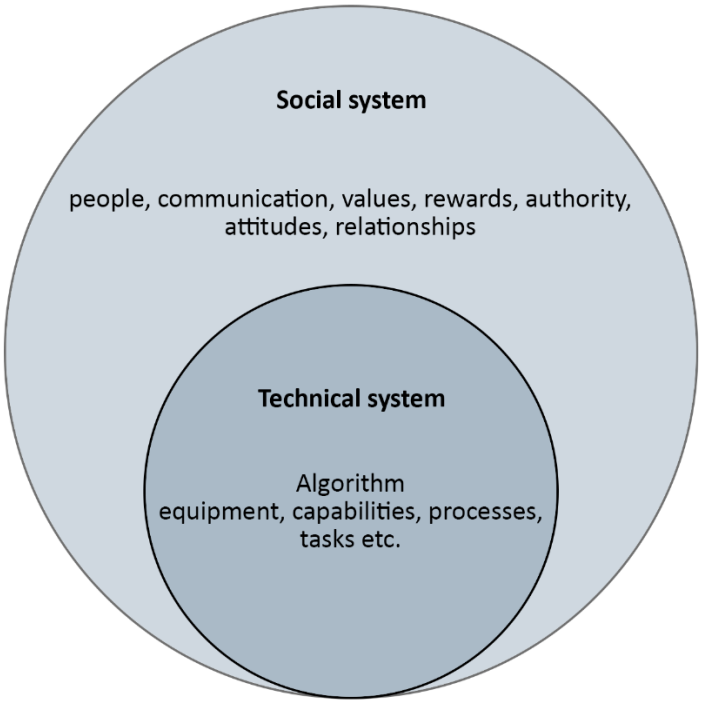


Figure 12 - Algorithm in sociotechnical system

Important to note here is that algorithms are part of a larger system. Fraud detection algorithms will not fully replace law enforcers, for example, the work beside each other. This is displayed in Figure 12 and might be referred to as a sociotechnical system. A sociotechnical system can be explained as 'design that consider human, social and organisational factors, as well as technical factors in the design of organisational systems' (Baxter and Sommerville 2011). This will be referred to as 'system'. In this specific context, the algorithm will be used as indication (fraud detection) for its social system, where humans and procedures (including investigation and verification) will continue. They work hand in hand. Another important point is that the algorithm and human experts act differently. This because law enforcers look at a client's data to detect possible fraud and they can immediately think about further research. Algorithms only detect fraud, but cannot further investigate. Therefore, the most interesting comparison is not necessarily that of human versus algorithm, but more that of human versus the system.

To find out what the differences between humans and AI are for the fraud domain, first the approach of an algorithm is investigated. An anonymous dataset had been created by Gemeente Amersfoort. From this dataset, Totta Data Lab created an interpretable dataset

for the algorithm. This interpretable dataset will be recreated by the author. After this, different (types of) algorithms will be tested to find the best fitting one. The best algorithm should give insight into the reasoning and results of algorithms. After this, interviews will be held with two of the law enforcers to discover the human approach. The results can then be analysed and compared to find the differences and similarities. This comparison will be made in the conclusion of this chapter.

5.2 Data understanding and preparation

Data understanding and data preparation are the second and third step in CRISP-DM, an open standard for data science projects. First, the data must be understood before being able to work with it. The database was provided in Microsoft Access[33] and comprised16 coupled tables. All tables were coupled on the *Clientnr* Unique Identifier (UID), which was pseudonymised. Totta Data Lab has sent the tables and fields that they had received as well, and the author had access to the exact same data as Totta Data Lab. They used 15 tables of the 16 tables. Four fields that Totta Data Lab used were not in the data set the author received, but three of these four could be reproduced. There were 5 fields in the dataset that were not in the dataset of Totta Data Lab, and these have been removed to make it as much alike as possible. In the end, one field was missing in the author's dataset compared to the data set Totta Data Lab used.

From the documentation provided by Totta Data Lab on Gemeente Nissewaard (Totta Data Lab 2020), it became clear that all cases where the social security assistance was either ended or when less money was provided were counted as fraud. This does not mean that the person did it on purpose; it might be that the municipality accepted the social security assistance, while this should not have been the case. Gemeente Amersfoort, however, does not agree on this. They count fraud when the client got their assistance stopped, reclaimed, suspended, or when their assistance is reclaimed and they are signed up at the labour inspection. The latter is also used for this research, since those results are in the dataset. This might lead to a less accurate algorithm, since there are fewer cases to train with, but it does lead to a more tailored algorithm to the organisation.

[33] https://www.microsoft.com/en/microsoft-365/access

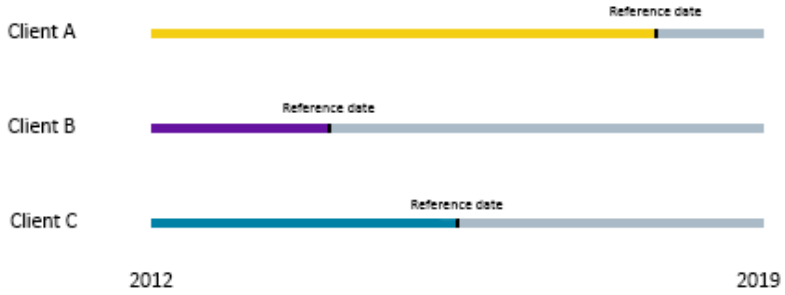


Figure 13 - Reference date per client, coloured parts are data that are taken into account, grey parts are data that are being discarded

For each client individually, a *reference date* is randomly generated. Then all data up until then is requested or aggregated. For example, if the reference date for a client is 14th of March 2017 and they birthed a baby in 2018, the number of children would not include this child. A graphical representation of this reference date can be found in Figure 13. This *reference date* is created to tackle the problem of the data set only going 7 years back and that of different levels of completeness of client files. Files of people who committed fraud or who were suspect of fraud were more complete, because extra checks have been performed in which extra information was gathered. This would make an algorithm able to 'predict' fraud cases based on completeness of a file; a more complete file equals a higher chance of fraud. In the real-world, however, this is not a good predictor, because fraudsters are not yet detected here, so the files of fraudsters are also not yet more complete. Therefore, an algorithm must be trained with incomplete data, because this imitates the real-world. To do this, the reference date has been introduced.

Then the dataset had to be altered such that there is only one row per client. This because all cases (clients) are classified by the algorithm into the class 'fraud' or 'no fraud'. These classes are thus predictions of the algorithm. Classification is done by the algorithm over one row (feature vector) with different attributes per client only. An attribute can be a feature directly, but if there are multiple rows of the same attributes for one client, these need to be aggregated in some way to be used by an algorithm. Multiple rows per attribute per client indeed occurred in the dataset, for example, with vacations. This meant that a client can be in a dataset of vacations twice, in two rows, simply because went on vacation twice. This should then be altered to an aggregate of the *number of vacations* per client. See Figure 14 for a graphical representation.

Clientnr	*Vacation Start date*
1	1-10
2	4-2
2	5-7
3	11-7

Clientnr	*# Vacations*
1	1
2	2
3	1

Figure 14 - Altering of data set in case of double rows by aggregation

Some tables consisted of multiple rows per client, while not being data that could simply be aggregated. In these cases, the last occurrence before the reference date was picked. In case multiple last occurrences happened on the same date, two things could happen. Either one date was picked at random when only a few double values occurred. Or, when many double values occurred, they were turned into Booleans. For example, see Figure 15.

Client nr	*Days away*	*Activity*
1	52	1
1	52	3
1	52	4
1	167	3

Client nr	*Last activity*	*Activity 1*	*Activity 2*	*Activity 3*	*Activity 4*
1	52	1	0	2	1

Figure 15 - Altering of data set in case of double rows by creating Booleans

In case of missing values, there were three options. First, these could be categorical values, in which case missing values were replaced with a new category, 'unavailable'. Second, when they were aggregates, they were replaced with a 0. Example of this is when a client was not present in budget checks, it means that their budget simply was not checked. Hence, the number of budget checks for that client can be set to 0. Third, when mainly the extremes were important, or it was a numerical value, and the first two options could not be applied, the median would be taken. Example of this is the age of the youngest and oldest child. This was all done with the library Caret in R.

Four additional changes were made to the data to make it interpretable for the algorithm. First of all, all date fields are changed to numeric fields (i.e., days away from the reference date), to give the algorithm an idea of timespan, which it would not get with only a date field. Second, *birth dates*, of both clients and children, were changed to age in years (according to the reference date). Third, when a client had multiple children, it was difficult to give their ages to the algorithm as well. This has been done by making a *min age* and *max age* field, for the youngest and the oldest child respectively. Fourth, all fields that directly indicate established fraud were left out. Fields that were discriminating or ethically questionable, were left out as well. These were either mentioned by Gemeente Amersfoort (*municipality client is living in*) or something the author felt highly uncomfortable working with (*hospital visits and death of children*).

After all this, an extra field was added which indicated whether someone committed fraud (1), was checked but did not commit fraud (0) or whether it was unknown (2). In the end, only people who were fraudsters (1) or who were confirmed innocents (0), were used for the algorithm, and the unknowns (2) were removed, because the most certainty about these cases was present. This reduced the dataset from 15181 entries to 1093 entries.

The attributes used in all tables can be found in Appendix F

5.3 Understanding the algorithm

Creating a simple decision tree will be very informative to provide insights in the way of working of algorithms. A simple decision tree was chosen because they are easy to train and to understand. Decision trees also make for easy visualisations, therefore easily communicable. Furthermore, decision trees compare to how humans may take decisions, in a simplified way, therefore seem appropriate for these data and this problem. A major drawback of simple decision trees are their, relative to other, more advanced algorithms, low accuracy. Reaching the highest accuracy here is not the most important aspect, since the algorithm mainly needs to be understood for comparison purposes with the human experts. An example decision tree is given in Figure 16.

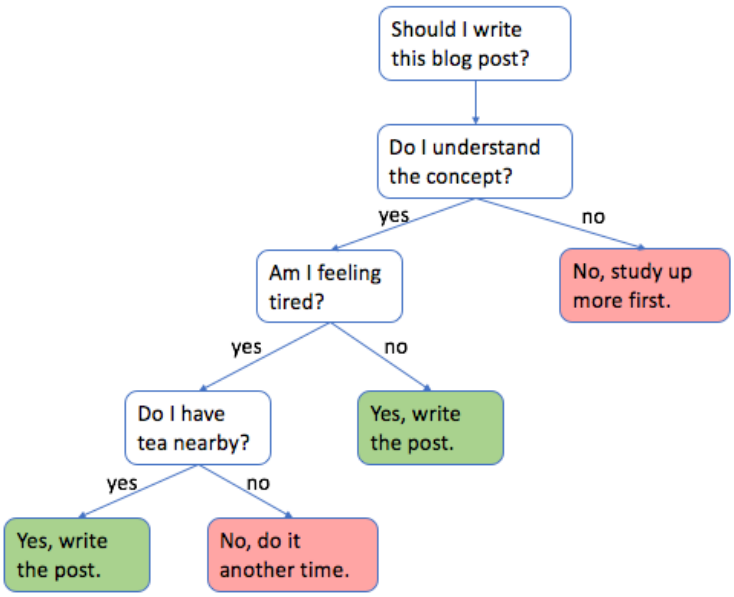


Figure 16 - Example of a decision tree [source: https://www.jeremyjordan.me/decision-trees/]

To create this algorithm, the R libraries Caret and RPlot are used, because of their ease of use. First of all, a seed is set. This seed is there so others can reproduce the results by a simple re-execution of the code. The data was then split into two sets. One for training and one for testing. This splitting is done randomly. The training data set comprises 70% of the data and the testing data set of the remaining 30%. The training data set is trained to fit a model. After this is done, the model is applied to the test data set to see how well it performs. The decisions are visualised so it can be easily seen which decisions were made. There are many methods to make the decision tree. Several algorithms will be tested so it can be seen what the 'best' model will be, for this specific case. It can still be useful to study the overlap between columns chosen by the algorithm produced by the different methods. It is also good to realise what is considered a 'best' algorithm, since this does not necessarily need to be the algorithm with the highest accuracy. The 'best' algorithm might very well be an algorithm that trades some of the accuracy for one that results in higher absolute number of fraudsters caught. One of the goals of Gemeente Amersfoort is to talk to every client at

least once a year, so, depending on how these conversations are being held, this might still be desirable.

5.3.1 Information split

First of all, a simple decision tree was created where the attributes were split based on what the algorithm found would gain it the most information. This algorithm favours subsets with many distinct values. The method of repeated cross validation was chosen (with 10 numbers and 3 repeats) for the trainControl method needed to train the algorithm. This resulted in an accuracy of 64.22%, which is about similar to that of Totta Data Lab. The confusion matrix for this method can be found in Table 2. This means that 13 real fraudster (true positives) were detected. 197 innocent people were marked innocent by the algorithm as well (true negatives). There were 101 fraudsters undetected by the algorithm (false negatives). Last, 16 people were marked as fraudster while they were actually innocent. This might have a large impact on their lives, since they (and their data) will be further research by human experts.

Table 2 - Confusion matrix for information split

Reference Prediction	No fraud	Fraud
No fraud	197	101
Fraud	16	13

Having achieved sufficient accuracy, more interesting is the decision tree itself, which can be found in Figure 17. What can be seen here is that the algorithm reasons that people with at least 1 *balance claim* (nl: *saldovordering*) and who had the last *debtors registration* (nl: *debiteuren registratie*) more than (or equal to) 1347 days ago are at risk of being a fraudster, as well as people with at least one *balance claim* (nl: *saldovordering*), who had their *last debtor registration* (nl: *debiteuren registratie*) less than 1347 days ago, their last *balance claim* (nl: *saldovordering*) less than 593 days ago and their *last balance check* (nl: *vermogenscheck*) more than 596 days ago. Underneath every final node, the number of correctly predicted customers is displayed in relation to the total observations in the training set. For example, of the 766 total observations in the training set (=70% of the total of 1093 observations), 33 have had their *last debtor registration* more than, or equal to, 1347 days ago and had at least 1 *balance claim*. Of those 33, 26 were indeed committing fraud. Interesting is that the algorithm uses only financial data of the customers. Apparently only these financial data were sufficient for the algorithm to reach the accuracy. It might be that the algorithm now checks on certain types of fraud, while other types go undetected. The algorithm is very simple and is not able to calculate complex relations between attributes and likelihood of fraud; it only looks at superficial determination.

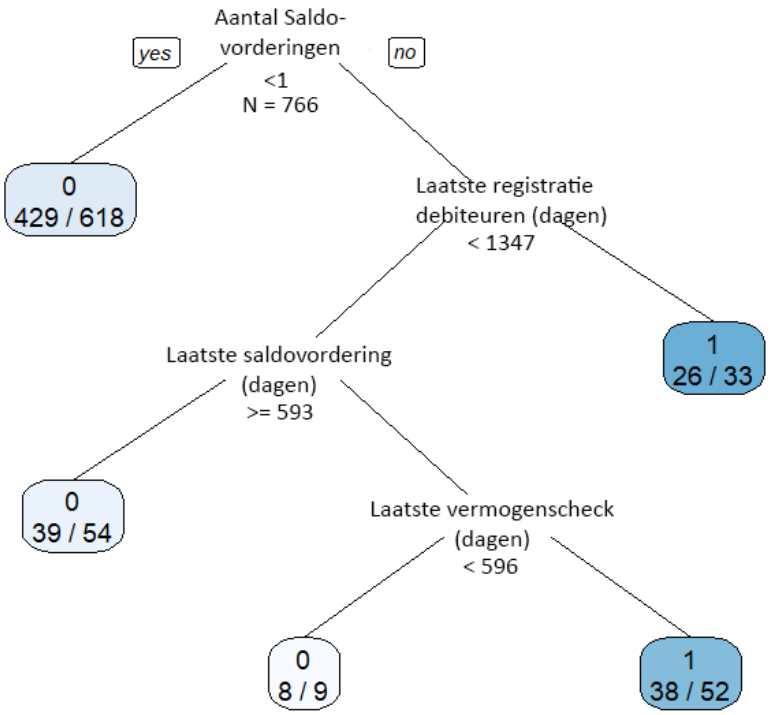


Figure 17 - Decision tree for information split

5.3.2 Gini index

Second, the Gini index has been chosen as a method of data splitting for a decision tree. This method looks for anomalies in the data and favours larger subsets. The method of repeated cross validation was chosen here as well. This algorithm had an accuracy of 63.91%, which is exactly one correctly predicted case less than the one on information split. The confusion matrix for this method can be found in Table 3. Although the accuracy being rather similar to the information split, there are more fraudsters being discovered by this decision tree at the expense of also more false positives. This makes sense, since there are two trees now, but they are still relatively simple trees.

Table 3 - Confusion matrix for Gini index

Reference Prediction	No fraud	Fraud
No fraud	188	93
Fraud	25	21

The full tree can be found in Figure 18. Note that the right side of the tree is similar to the one split on information gain. This algorithm has discovered that there is an extra tree with possibilities of fraud. This is present when a client has had 0 balance claims (nl: *saldovorderingen*), two or more *icosigs* (signal) and at least one *work acceptance* (nl: *werkaanvaarding*). A possibility for this could be that someone forgets to inform the municipality of their newly found job.

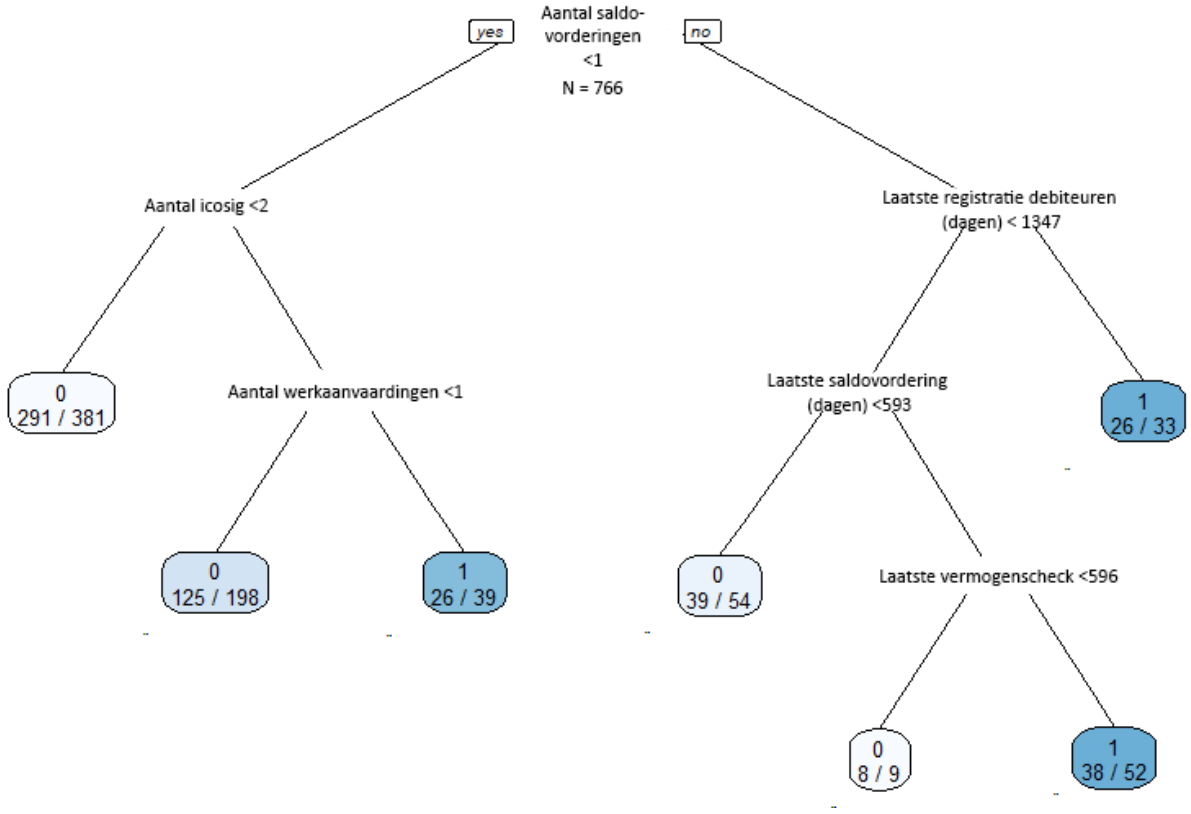


Figure 18 - Decision tree for Gini index split

5.3.3 Random Forest

In the technical documentation of Gemeente Nissewaard, Totta Data Lab mentioned having used Random Forest as a technique to create their algorithm (Totta Data Lab 2020). This will be reproduced below. First will be started with a brief explanation of what Random Forest (RF) entails, so the results can be interpreted better. The problem being tackled in this chapter is a classification problem, as mentioned earlier. One of the most popular classification algorithms is RF.

RF, contrary to what was done above, does not create one decision tree, but it creates many. The idea behind this is that a bias that might be present in one tree can be filtered out by complementing it with another, relatively uncorrelated, tree. It does not create a bias by constantly picking the same variables to make decisions upon. In the RF approach, decision trees are learnt for different subsets of the data and different subsets of the attributes. Combining the predictions of all trees by taking a majority vote, has proven to be a powerful tool for prediction. Although usually more powerful in making predications, the resulting

model (a set of trees) is less interpretable; it is not possible to generate as nice, visual tree, like before. What can be derived, however, is a list of most and least important attributes. How many attributes should be combined depends on the data and algorithm. In this case, around 28 attributes are the optimal value, as can be seen in Figure 19. This could be further optimised with more time and processing power available.

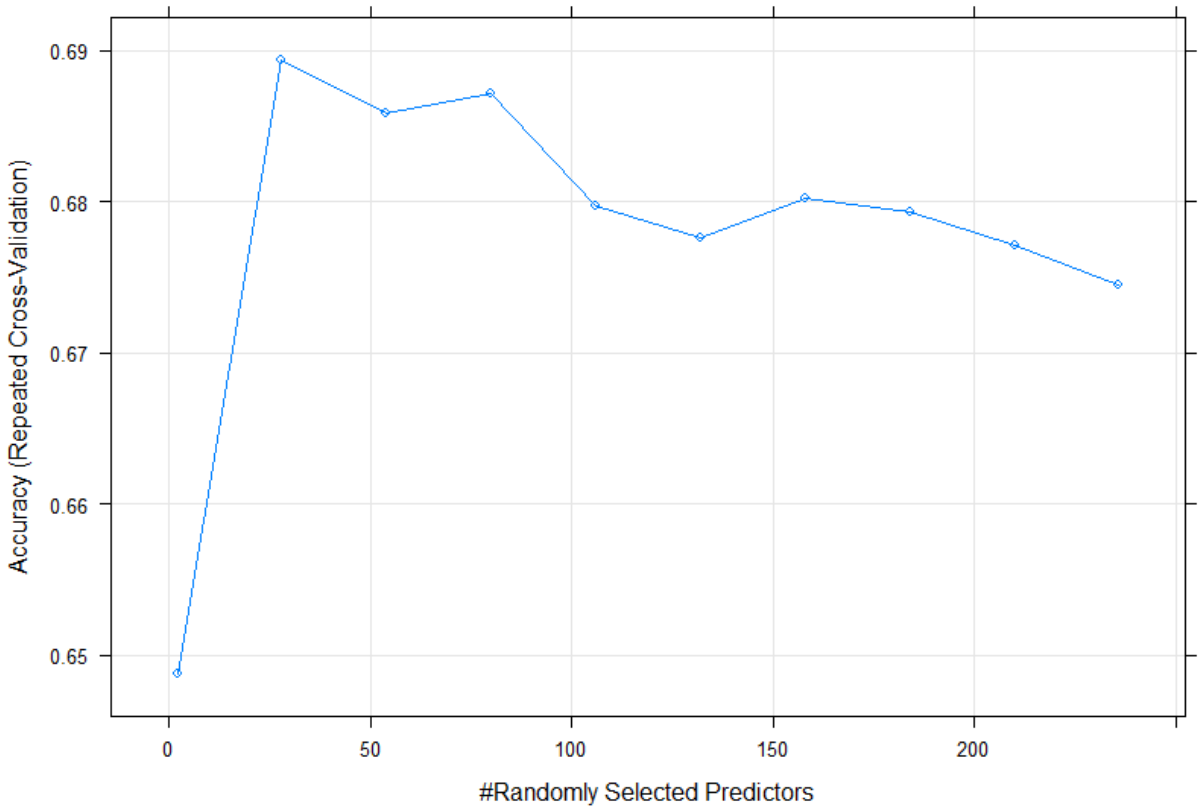


Figure 19 - Accuracy of the RF algorithm based on the number of predictors

One important part of RF is that the data is relatively uncorrelated. While this is, intuitively, not entirely the case with the dataset created, RF will still be used to replicate, to some extent, what Totta Data Lab did.

For this algorithm, the Caret library of R was used. The accuracy of the RF algorithm was 69.42%. It makes sense that this is higher than the previous two algorithms, because it is a more advanced algorithm. The full confusion matrix can be found in Table 4.

Table 4 - Confusion matrix of the RF algorithm

Reference Prediction	No fraud	Fraud
No fraud	195	82
Fraud	18	32

Table 5 - Top 20 most important variables of the RF algorithm

Variable	Importance
DD_LAATSTE_BLOKKADE	100.00
DD_BEGIN_LAATSTE_RPO	97.83
LEEFTIJD_PEIL	94.27
DD_TREDE	92.17
AANTAL_ICOSIG	88.20
DD_LAATSTE_TRAJECT_START	73.32
REGISTRATIEJAREN	72.37
DD_BEGIN_LAATSTE_ACTIVITEIT	67.19
MAX_LEEFTIJD_KIND	67.03
MIN_LEEFTIJD_KIND	66.95
AANTAL_BLOKKADES	66.73
DD_LAATSTE_SALDOVORDERING	64.77
DD_LAATSTE_DEB_REG	61.12
DD_LAATSTE_VERMOGENSCHECK	59.17
DD_LAATSTE_ICOSIG	58.57
AANTAL_DEB	55.28
DD_EIND_LAATSTE_ACTIVITEIT	51.60
AANTAL_TREDES	41.94
DD_EIND_LAATSTE_RPO	40.40
AANTAL_DOSSIERS	40.01

Table 5 displays the top 20 most important values. From here can be seen that the date of the *last block*, *date of the last RPO*, *age of a customer* and *date of the last step* were very important for the algorithm to make decisions upon. All have an importance of more than 90%. In total, more than 200 attributes were present, so the bottom attributes had very low importance. Interesting is that the top 20 attributes are not mainly financial data like was seen previously, however all attributes highlighted by the Information split algorithm are present in this top 20. The top 20 attributes are wide-spread in terms of origin tables (and therefore topics).

5.3.4 Lasso

In the technical documentation, Totta Data Lab talked about the Lasso algorithm and the Rulefit algorithm (Totta Data Lab 2020). First Lasso regression was tried, again with the Caret library on the Gini index. Lasso is a technique where the number of variables will be lowered, by only keeping the most important variables. The accuracy of the algorithm was 63.91%. Its confusion matrix can be found in Table 6, which is identical to the Gini index. The tree is also identical to the Gini index tree, and can thus be found in Figure 18.

Table 6 - Confusion matrix Lasso algorithm

Reference Prediction	No fraud	Fraud
No fraud	188	93
Fraud	25	21

5.3.5 Rulefit

Another algorithm mentioned by Totta Data Lab is a Rulefit algorithm. This algorithm can be created using Caret's method, C5.0.

The Rulefit algorithm creates many if-then rules. These rules are then tested and the best ones are kept. This seems similar to the simple decision trees mentioned earlier, however, Rulefit can also express interactions. Rulefit has the benefit of outperforming a simple decision tree, while being rather similar in performance to more complex algorithms like random forest. A benefit of Rulefit over RF is that it is better interpretable, since the rules given are hard and clear rules. An example of such a rule on this dataset is given below:

Rule 34/6: (12.5, lift 2.3)

AANTAL_TRAJ_TWL <= 0

AANTAL_BLOKKADES > 6

AANTAL_DETENTIE <= 0

AANTAL_AANVRAGEN > 0

-> class 1 [0.931]

C5.0 with Caret resulted in an accuracy of 67.58%. The confusion matrix can be found in Table 7. As can be seen here, there were more false positives present than in RF. This also made the accuracy lower. This might be because the algorithm is rather simple; it does not entail more than many simple if-then statements.

Table 7 - Confusion matrix Rulefit algorithm with Caret

Reference Prediction	No fraud	Fraud
No fraud	174	67
Fraud	39	47

The main attributes used by this algorithm can be found in Table 8. These are in no particular order. All rules could be displayed here, however there were way too many rules found by Rulefit to display here and some of them were confidential, since they could be identifying. Rulefit looks at many of the same attributes as RF did (*blocks*, *registration years*, *# debtors*), however it also specifically looks at certain categories within those attributes. In terms of approach it is similar to decision trees, because it is a rather simple and straightforward algorithm, and to RF, because it creates many such decision trees. This can also be seen in the resulting accuracy; it lies between the (simple) decisions trees and RF.

Table 8 - Attribute usage of Rulefit with Caret

Attribute usage:
REGISTRATIEJAREN
LAATSTE_DOS_KODE_REG=unavailable
DD_TREDE
CODE_TREDE=A1
AANTAL_PW_BIJZ
AANTAL_SALDOVORDERINGEN
DD_LAATSTE_SALDOVORDERING
LAATSTE_ACT_ACTST=02
LAATSTE_ACT_ACTIVITEIT=2WP
LAATSTE_ACT_ACTIVITEIT=INM
DD_LAATSTE_BLOKKADE
AANTAL_DETENTIE
AANTAL_EINDE_WIJ
AANTAL_AANGANG_RELATIE
AANTAL_GEEN_INLICHTINGEN
AANTAL_DEB
BEGIN_LAATSTE_BIJZSIT
AANTAL_ICOSIG

5.3.5 MARSplines

The last algorithm mentioned by Totta Data Lab was MARSplines. This stands for multivariate adaptive regression splines. MARSplines makes no underlying assumption on the relationships between the data. It creates this relation from the data. It might be that one attribute gets a higher weight than another variable. It thus assigns weights to certain attributes (like earlier algorithms did too, such as Lasso). An example of this is the cost of a car. The cost is likely depending on the age of a car. MARSplines can create the breaking points to detect this relation. Hence, the cost for a car that is over 10 years old would be lower than the cost for a car between 0-10 years old. Hence it finds the most logical patterns. The results of this algorithm can most easily be interpreted by displaying the most important attributes, like by RF.

MARSplines can be created by using Caret. The accuracy found by the MARSplines algorithm was 67.89%. The confusion matrix can be found in Table 9.

Table 9 - Confusion matrix MARSplines

Reference Prediction	No fraud	Fraud
No fraud	183	75
Fraud	30	39

The most important attributes can be found in Table 10. As can be seen, very few attributes were chosen eventually. The most important attributes are similar to the used attributes of the simple decision trees in 5.3.1 and 5.3.2.

Table 10 - Most important variables MARSplines

Variable	Importance
AANTAL_BLOKKADES	100.00
AANTAL_ICOSIG	89.87
LAATSTE_SALDOVORDERING	81.63
LAATSTE_ACT_ACTIVITEIT=2WP	62.32
BEGIN_LAATSTE_RPO	54.37
DD_TREDE	47.49
MAX_LEEFTIJD_KIND	39.12
AANTAL_DETENTIE	31.97
AANTAL_WEEKBETALINGEN	18.85

5.3.6 Conclusion of the algorithm's results and reasoning

It is good to emphasise here that the goal of this chapter is not to create a perfect algorithm. Even though the accuracy is sufficient, there are endless methods that might increase accuracy or that produce less false positives specifically. It has to be kept in mind, though, that the ultimate goal of this research is to understand how a learned algorithm reasons to be able to compare it to how human experts reason. The algorithms that are tested are the ones either mentioned by Totta Data Lab (Totta Data Lab 2020), or ones that provide this research with a better understanding of the outputs of the algorithms.

Another thing which is good to mention here is that the prediction task is challenging and the data rather complex. Therefore, it is not expected that the algorithm works with an almost perfect predictability. One reason why the prediction task is so challenging, is that it needs to be able to predict at any point in a customer's process, hence also when not much interactions have happened yet and much data is still missing (artificially introduced into the training data with the reference date technique). Next to that, there are many fields included in the algorithm, even more for algorithms that cannot deal with categorical data such as RF (in these cases, a categorical variable is transformed into several Boolean variables indicating "variable has value X", creating a larger, more complex data set which requires more processing power). The accuracy found above (~67-69%) can be considered rather high, considering the complexity. Again, with more advanced methods and with more

processing power, accuracy can probably be increased, but that is not the purpose here. Moreover, more advanced methods typically produce less interpretable models, hence we cannot reflect on how they reason.

Concluding from the algorithms that were tested above. First to acknowledge is that all the algorithms that were tested had a higher accuracy than picking randomly. All algorithms had an accuracy of around 65%, whereas randomly picking would give an accuracy of 50%. The dataset that was used is skewed towards 'no fraud'. 65% of the clients in the dataset did not fraud (213/327). Hence, an algorithm could reach a similar accuracy to the algorithms tested above by simply always guessing 'no fraud'. Then no clients would be further investigated and the algorithm would not be of any value. Picking randomly, on the other hand, would make it such that many people would be marked as fraudster while they were not (false positives). This is also not desirable. The dataset with which was tested only included people who were checked, which resulted in 35% fraudsters and 65% confirmed non-fraudsters. In the full dataset this percentage of fraudsters will be much lower (around 2.5%), which might influence the accuracy.

In the technical documentation provided by Totta Data Lab on the algorithm of Gemeente Nissewaard (Totta Data Lab 2020), they found that Random Forest (RF), combined with MARSplines gave the best results. Even though their dataset and algorithm will differ from the one used during this research, this research indicated similar results, as can be seen in Table 11. Unfortunately, Totta has not officially openly communicated their exact accuracy, especially not in this case. Hence, only the best algorithms from this research can be compared. After training, all algorithms could classify the entire testing dataset in a matter of seconds.

Table 11 - Accuracy of the different algorithms

Random forest	MARSplines	Rulefit	Information split	Gini index	Lasso
69.42%	67.89%	67.58%	64.22%	63.91%	63.91%

The accuracy itself is not the only parameter to look at when deciding upon a good algorithm. The confusion matrices are too. Especially during the ethical part of this research, it was found that being a false positive has quite a large impact on a subject. Hence, this should be prevented as much as possible. However, the fraudsters still need to be caught and undoubtedly innocent people will be marked in this process as well. Being a false positive has a larger impact on an individual than being a false negative. However, society benefits from catching as many fraudsters, hence wants not too many false negatives. It should be noted that one of the goals of Gemeente Amersfoort is to talk to every client at least once a year. Depending on how these conversations are being held, it might be a bit less impactful on someone when they are a false positive, since they expect a conversation anyway. There should still be kept in mind that it does have a large impact on an individual when they are marked as a potential fraudster. When looking at the percentage of false positives versus true positives, RF also scored the best. Of the fraud cases marked by the algorithm, 64% actually frauded. All other tested algorithms scored around 55%. Both RF and

MARSplines seem to have good performance in both accuracy and the (low) relative number of false positives. Good usage of this could, for example, be to combine the two algorithms, like Totta Data Lab mentioned, or to use the more accurate one first (RF) and later on add the other one (MARSplines), so they can complement each other. Noticeable is that Rulefit also performed quite well, so that algorithm might be used as well, depending on preferences. In this research, the RF algorithm will be used for simplicity reasons and because it attributes the most to the final model of Totta Data Lab (Totta Data Lab 2020).

The most important results from comparing the algorithms are on how the algorithms work, such that this can be compared to how humans would do this. For the simple algorithms, such as the decision tree, it was possible to visualise how they exactly work. However, for the more complex algorithms, this was not. For these, the variable importance has been displayed. The variables have been grouped based on their origin table. The most important tables overall are displayed in Table 12.

Table 12 - Overview of tables and variables within those tables used per algorithm

Table	Variables per table used within all algorithms (/total variables per table)	Information gain	Gini	Random forest	Lasso	Rulefit	MARSplines	Table used by # algorithms
Debiteuren (debtors)	5 (/12)	3	3	3	3	2	1	6
Blokkades (blocks)	8 (/16)		1	2	1	5	3	5
Icosig (incoming signal)	2 (/2)		1	2	1	1	1	5
Vermogen (capital)	1 (/3)	1	1	1	1			4
Activiteiten (activities)	5 (/88)			2		3	1	3
Kind (children)	2 (/3)			2		3	1	3
Trede (steps)	3 (/24)			2			1	2
Dossiers (dossiers)	2 (/22)			1		1		2
RPO (RPO)	2 (/11)			2			1	2
Cliënt (client	2 (/5)			2		1		2
Participatietraject (participation process)	1 (/15)			1				1
Bijzondere situatie (special situation)	1 (/8)					1		1
Vakantie (vacation)	0 (/5)							0
Contacten (contact moments)	0 (/15)							0
Aanvragen (requests)	0 (/7)							0
Total		4	6	20 (213)	6	17	9	

Noticeable here is that random forest uses the most variables and information gain the least. This is logical since those algorithms are one of the most complex and one of the simplest, respectively. Furthermore, all algorithms used variables of the debtors table. Almost all algorithms, except for the very simple information gain decision tree, used variables from blocks and incoming signals. One last thing should be mentioned here, which is that RF uses way more variable than the 20 displayed, namely 213 (of a total of 236). Only the top 20 were taken for this part of the research. When looking generally at RF, this algorithm looks at relations of (combinations of) attributes to find a common link between data and committing fraud. This also goes for combinations of attributes that might not feel logical for humans to compare. A general remark on the usage of an algorithm is that, after training, the 327 could be classified in a matter of seconds, while humans would take way longer.

Specifically, for the three best performing algorithms (RF, MARSplines and Rulefit), including the ones mentioned to be used by Totta Data Lab, five other observations can be made. First, Rulefit especially uses *blocks* and within *blocks* it looks for special categories, like number of times in jail (*aantal_detentie*), number of times WIJ (law investing in young people) was ended (*aantal_einde_WIJ*), number of times someone started a relationship (*aantal_aangang_relatie*) and number of times the municipality was not informed (*aantal_geen_inlichtingen*). Second, Rulefit uses the tables activities, debtors and children (kind) mainly. Third, as mentioned, random forest uses many attributes from many different tables. Blocks, activities, steps, debtors, RPO, client and incoming signal are the tables mainly used. Fourth, tables special situation is barely used by any of the algorithms, as well as participation process and dossiers. The tables about vacations and contact moments are not used as top predictive value by any of the algorithms. Last, RPO and client are important for RF, but not very important for the other algorithms.

What should also be noticed is that any algorithm can only use data provided to it; algorithms cannot enrich these data. Algorithms also do not have the ability to have empathy. They look at the data as-is, without taking the human aspect into account. For example, the algorithm does not know what detention entails, whereas humans do. Therefore, algorithms do not have (unconscious) prejudice, whereas humans may. This also works the other way around, when human experts would clear someone from being a fraudster because of prejudice.

Something often mentioned by different stakeholders and by the lawfulness part of this research is that it would be nice to take the algorithm which needs the least variables. One such algorithm in this case would be the MARSplines algorithm, which only needs 9 variables to reach an above-average accuracy. While on the one hand this would indeed be good, this would also create room for more bias, since the algorithm would rule out the other variables. Hence this does not need to be an argument for MARSplines or against RF or Rulefit.

In short, the algorithms with the most potential seem to be RF and MARSplines, as mentioned by Totta Data Lab as well, whereas Rulefit seems promising too. These were chosen because of their accuracy and the relative low number of false positives. RF looks at correlations of attributes to detect patterns. These patterns may also originate between combinations of attributes that may not feel logical. The goal of this part was not to create the best algorithm, but to understand these algorithms for comparison purposes later on. Blocks and incoming signals seem to be the most used tables by the algorithms created. Special situation, dossiers and participation process are the least used, and vacations and contact moments are not used at all. In general, all algorithms preferred using financial data. Interesting was that some algorithms also looked for certain categories within tables, like if someone was in jail. Furthermore, algorithms were very fast; humans could by no means match this speed. The algorithms could only use the data provided to them for the defined goal without being able to enrich this data or alter the goal (by, for example, also helping customers). Algorithms did not have prejudice about the data, the algorithm does not even know what certain attributes entail. Last, less variables might not necessarily be better, since this might introduce bias.

5.4 Understanding the human experts

The results from the conclusion above must be compared to the reasoning and results of the law enforcers to be able to see the difference, and to also see the difference between the system. To do this, interviews were planned with two law enforcers. These interviews were again held in a semi-structured manner, as was chosen during earlier interviews too. During the interviews, it was asked what they feel are the most important criteria to consider someone possibly guilty or possibly innocent. Next to that, they were asked to rank ten cases from most possibly guilty to most possibly innocent. To be able to ask the latter question, first a dataset must be prepared.

5.4.1 Data preparation

The dataset that was prepared for the law enforcers consists of 10 cases. To be able to really compare the algorithm to the law enforcers, their task should be as similar as possible to that of the algorithm. The cases also include false positives and false negatives, as were found by the RF algorithm earlier. In the dataset for the law enforcers, the ratio of fraud cases and non-fraud cases is equal to the dataset of the algorithm. Furthermore, three cases that were misinterpreted by the algorithm were included; one false positive and two false negatives. These cases were picked randomly. It would be expected that the false positive and the false negatives would end up somewhere in the middle of the ranking, because cases that were challenging for the algorithm will probably also be challenging for the human experts. All data that was available about these 10 people has then be collected from the original dataset. Data that would directly indicate whether someone is innocent or not was removed, and data that might directly identify certain people were aggregated (dates were changed to years). All in all, the law enforcers will get a larger, more extended, dataset than the algorithm. This is mainly because humans are able to interpret these data fields better, because of assumptions and experience. For an algorithm to be able to understand data, the data must be prepared a lot more.

5.4.2 Interviews

The interviews were held in the second half of December 2020. Unfortunately, these interviews had to be held via Microsoft Teams, instead of in person, due to COVID-19. This entailed that the challenges of Microsoft Teams and general computer issues had to be dealt with. This influenced the interviews. For example, not all tables were taken into account by one interviewee, because Excel would crash on the interviewee's computer if they tried to. A second point to be made here is that ranking the clients was quite difficult, rather they were categorised in 'to research', 'to keep an eye on (or 'may be fraud')' and 'nothing special noticed here'.

First, the results about the criteria that would be taken into account by the law enforcer to decide whether someone is possibly innocent or guilty. The first interviewee mentioned that, when a signal comes in, they look at the dossier of the client to see, amongst others, how long they are already receiving social security assistance, how they perform in their participation process and at their bank statements. This is mainly to check whether or not the customer is in the picture. The second interviewee agreed with the money statement, and added that water usage might be interesting too. The first interviewee mentioned that the money related data gave them the most insights into whether someone might be guilty, while observations, conversation with the customer and house visits were the best indicators for someone being innocent. They both mentioned that it happens that many, similar, signals can come in about one customer from the same source. This might be an argument between the customer and the source. This is, however, a good indicator that someone is innocent.

Another thing that was mentioned by the both interviewees was that they spend quite some time on social media. They do this to verify stories they heard. For example, if the signal is that someone has a job, but did not declare this, they search for the employer to see if there is evidence the customer indeed works there. Another example is for someone illegally living in. This has to be stated at the municipality. If this has not been done, but there is suspicion, often social media evidence can be found of this. This was the clearest example of human experts doing both fraud detection and further investigation.

The main thing both interviewees mentioned that they value the conversation with the customer quite heavily. From this conversation, much information is gained. Customers are asked questions, and often it is rather clear whether someone can give a logical answer to those or not. The second interviewee especially mentioned that observations were very valuable to them.

After the discussion about the criteria, they were presented with the dataset. The tables that were most looked at by the interviewees were client, activities, special situation, blocks, debtors, children, RPO, participation process, vacations and capital. Icosig, steps and dossiers were not or barely used. It should be considered that this might be because they are used to another way of working, rather than Excel. The second interviewee mentioned not to use the Icosig table, because they did not want to be biased. The main decision whether someone was suspicious seemed to come from the table blocks, participation projects and

debtors. There was quite a difference between the interviewees. While both tried to profile their customers, one did this with fewer tables than the other. The second interviewee looked at the progress of the customer, while the first interviewee focused more on the data itself and whether that fit in the created profile. Furthermore, there was a difference when a different type of fraud was suspected. For example, vacation became more interesting when capital fraud was suspected.

Both interviewees made profiles of people with the data provided. For example, a certain customer had taken quite some vacations, so it was highly likely that they were not from Dutch origin. This was done to match the data known about them with their logical behaviour. Someone with lots of vacations may once have completed an integration course. Or someone with multiple partners may have had many changes in family composition. If something occurred here which was out of the ordinary, this was something to suspect. The interviewees did not really look for specific activities. Certain activities or reasons for blocks, for example, did stand out to the law enforcers, however. Furthermore, it was looked at whether the customer's progression is logical. An example of non-logical progression was someone who had been receiving social security assistance for many years and was, according to the data, not motivated to work. This might be legit, but chances are that someone becomes bored after a long time of not working. They might then undertake activities, like babysitting for their grandchildren. This might be a way in to get them back to work. Interesting here too was that the law enforcers looked for different data depending on the type of possible fraud they suspected.

Interesting was that interviewee 1 also sometimes mentioned that they thought someone was not suspected, but it might be good to still be in contact with them, because it had been a while or because it was questioned how well they were being guided. Interviewee 2 treated the dataset more like all 10 cases were incoming signals. For all ten cases, questions were thought of to focus on during a conversation with said customer. Some customers were less suspicious than others, but then it was mentioned that if they would commit fraud, it would probably be on living situation, for example. Here the sociotechnical system can be seen in place as well.

Next to this, what was interesting was that there seemed to be empathy. There was one case of a customer who committed fraud, but the interviewee mentioned that they thought the customer to be innocent, because the customer was almost going to retire. Again, this is part of the sociotechnical system, where humans do more than only detecting fraud.

Table 13 - Results of prediction of fraud cases of the law enforcers

	True positive (fraud)	True negative (innocent)	False positive	False negative
Correctly predicted by law enforcers	75%	75%	100%	0%

Finally, it was thought that false negatives and false positives would be difficult for the law enforcers too. The results can be seen in Table 13. For example, the false positive as marked by the algorithm was detected as negative (no fraud) by both law enforcers. The law enforcers were, logically, more cautious with marking someone as potential fraudster than the algorithm.

5.4.3 Conclusion of the human experts' results and reasoning

Concluding, the interviewees found financial data to be very interesting in case of an incoming signal. Furthermore, conversations and contact with the customer are highly valued. They mentioned repeated incoming signals as something which tells them someone might be innocent. Both interviewees agreed that social media can help in fraud research. When given the dataset, the interviewees used many tables to find out more about a customer. Tables that were not or barely used were Icosig, steps and dossiers. Blocks, participation projects and debtors seemed to be the most deciding tables. Both interviewees tried to create some context around the data provided to picture logical behaviour, and, in this way, find the odd ones out. Next to this, empathy seemed to be present with the law enforcers. Last, the law enforcers were more cautious in making decisions than the algorithm.

5.5 Conclusion

Figure 20 - Differences between the law enforcers and the algorithm

To answer the question 'In what ways is algorithm-based fraud detection different from human expert-based fraud detection?', an algorithm has been created to compare its reasoning and results with human experts' reasoning and results. It is important to realise that the algorithm implementation does not mean that the law enforcers will be replaced. The algorithm will help the law enforcers in better dividing their limited capacity. The algorithm does not make decisions, it simply signals some clients to the law enforcers that it suspects. Therefore, it is part of the sociotechnical system. This conclusion is summarised in Figure 20.

Five observations were made when comparing the methods of the algorithm and the law enforcers. This does not conclude one of the first differences noticed in the earlier chapters: the resources. These are limited for the law enforcer, while an algorithm has almost unlimited capacity. First, in terms of specific tables that were or were not used, it was noticeable that both the algorithm and the law enforcers seemed to have quite some interest in the financial data from the customers. Furthermore, it stands out that blocks was

a very important table to both the algorithm and the human experts. Icosig was important to the algorithm, but not to the law enforcers. This was partly because the law enforcers thought to interfere with the experiment if they looked at it. This might be because an Icosig is created once someone is a fraud suspect. Because of the reference data that was introduced, this problem was tackled. Interesting is that special situation and participation process were barely used by the algorithm, while they were very useful to the law enforcers, to create a better picture about the customer. The same goes for vacations, which was one of the deciding factors to research someone or not for the law enforcers. Dossiers and contact moments were not used by both the humans and the algorithm. When the system (law enforcer + algorithm) will work as a whole, more data will be used than by either the law enforcer or the algorithm.

Second, as briefly mentioned above, the law enforcers tried to create a profile for each customer. This to understand what the customer's logical behaviour would be. An algorithm also looks at patterns. The law enforcers, however, have more background knowledge about the social security assistance process and the meaning of certain data (which differ per type of suspected fraud). Both ways of working have their own benefits and drawbacks. An interesting change to the algorithm, based on the interviews, would be to choose (and test) an algorithm that also creates profiles, like the law enforcers did too. This is, however, looking from a data point of view, since ethically (and legally) some cautiousness must be taken. When the system as a whole is in use, the law enforcers will still create personas, but these will be created for the risk cases, marked by the algorithm, only. Note that outside this system, the regular signals still come in, like described in the first sub question, but these are not considered for this part.

Third, interesting is to see that more social tables are used by the human agents to paint a better picture about the client, while the algorithm does not paint such a picture. Law enforcers can also enrich this data with, for example, bank statements, Kadaster data or conversations, whereas the algorithm cannot. This is part of the system, where further investigation is performed. Social media data, however, might already be used in the fraud detection part (before the first green light moment). Hence, the law enforcers seem to create a more imaginable picture to find logical and illogical patterns. This might partly be the case because they want to help the customer, by, for example, talking to their customer manager, rather than only catching fraudsters. This fits the observation that they were more cautious to mark someone as a potential fraudster than the algorithm. An algorithm does not care for this. Again, using the system as a whole will not mean that the law enforcers cannot be cautious or helping a customer anymore, since the algorithm will complement the law enforcers. The algorithm does the fraud detection via risk indication, where the human experts can validate this risk indication and further enhance the data. In this second phase, the law enforcers can also opt to help people and let the investigation as-is, for example, or give people the benefit of the doubt. Hence, in the end, the law enforcers make the decisions, which can still be cautious, the algorithm merely highlights customers it is suspicious about. In addition to this, Gemeente Amersfoort is looking at the creation of a more social algorithm as well, to see whether they can detect people who may need to receive more or different assistance, which seems like an interesting way to go.

Fourth, there were some differences in method found between the law enforcers. While there were differences between different algorithms, once an algorithm is chosen, these differences are not present anymore. For the law enforcers, these differences will keep being present because they are human. Even when introducing an algorithm, this will still occur, since the law enforcers will have the final say. Not only were there differences in approach per law enforcer, there were also differences for different types of fraud. This did not show in the data that were used, but in the preparation for the conversation with the client it did. Different questions would be asked or prepared to be asked for different types of fraud. Hence it might be interesting to create multiple algorithms for multiple types of fraud, because different criteria might become more valuable. A well developed learning algorithm will detect different types of fraud, but it might be easier to understand the algorithm's reasoning when different algorithms are created for different types of fraud and these algorithms might be tailored more to a more specific type of fraud.

Fifth, the law enforcers had some empathy for the customers. Empathy is not present at all for the algorithm, so on this part an algorithm will probably be more objective. Important to keep in mind here, is that the empathy will not be gone when introducing the algorithm. This is because the algorithm of Totta Data Lab will be an addition to the law enforcers, rather than a full replacement. Another important point is that this might differ in other municipalities or organisations, since Gemeente Amersfoort mentioned that they are a more social municipality than others.

Chapter 6

6. Ethical Framework

In this chapter, an ethical framework for algorithmic fraud detection at municipalities is created. This framework is validated afterwards. The short version of the framework can be found in Figure 21, the extended version in Appendix C

6.1 Method

In this chapter, a short framework will be created based on the related work. This short framework will be extended, to make it more workable in practice, but also to highlight the important aspects per topic of the framework. This may be used to get background knowledge on certain parts of the framework or as reference for others. A brainstorm will be performed with professionals in this area to also include non-literature sources and to include different angles. In the end, the created framework will be validated using two practical case studies. With this, the question 'Which ethical considerations should be taken into account when using a fraud detection algorithm?' should be answerable.

6.2 Brainstorm[BR]

The ethical frameworks found in related work presented in 2.7 were combined to create an initial framework. This framework was used as a discussion starter with various professionals from both the research area and the working force. From these brainstorms, three main changes were made. First, the addition that the algorithm must be fair was made was suggested from a brainstorm with Maurice van Keulen, supervisor of this research. Earlier, this was placed under 'the algorithm respects the rights and interests of all parties impacted', but it was found that this was not sufficiently explicit. Since the point is a rather important one, it was decided to state it explicitly. Second, everywhere where now states 'the algorithm', then stated 'the AI'. It was mentioned by Maranke Wieringa, doctoral candidate at Utrecht University on a similar research field, that AI may sound too advanced. While actually similar ethical principles apply to having an algorithmic function in, for example, Microsoft Excel. This might have caused some people to not use the framework because of thinking their application does not qualify as AI. Therefore, the choice was made to change the semantic to 'the algorithm' instead. Third, semantics were added to the framework. These were mainly suggested by one of the experts brainstormed with, Maurice van Keulen. For example, 'as little privacy infringement of people' was changed into 'as little and justifiable privacy infringement of people'. With these changes, the framework was finalized.

6.3 Project framework

It is useful for this project to create a framework for the ethical side which can be used by municipalities for the ethical side of algorithms. For this purpose, mainly the AI4people framework of 2.7.1 will be followed. Although some small additions will be made to include the results from the related work and important points of other frameworks. Next to additions, also some extra substance in terms of bullet points is given to the framework, to make it more workable in practice. This framework with the additions can be found in Figure 21 below. The additions are in italics, to make it more easily interpretable. The additions have a small caption at the end, to indicate where they come from, where 2.7.2, 2.7.3, 2.7.4, 2.7.5, and 2.7.6 are from the other ethical frameworks in corresponding chapters and BR is from the brainstorms. In the content, also the letter of the ministry of justice and security is taken into account, since this letter will probably lead to new guidelines from the government. In the text, the main focus will lie on the fraud detection algorithm as this report is as well. The framework will be validated by applying it to cases from practice, like SyRI and the algorithm of Gemeente Amersfoort and by consulting professionals.

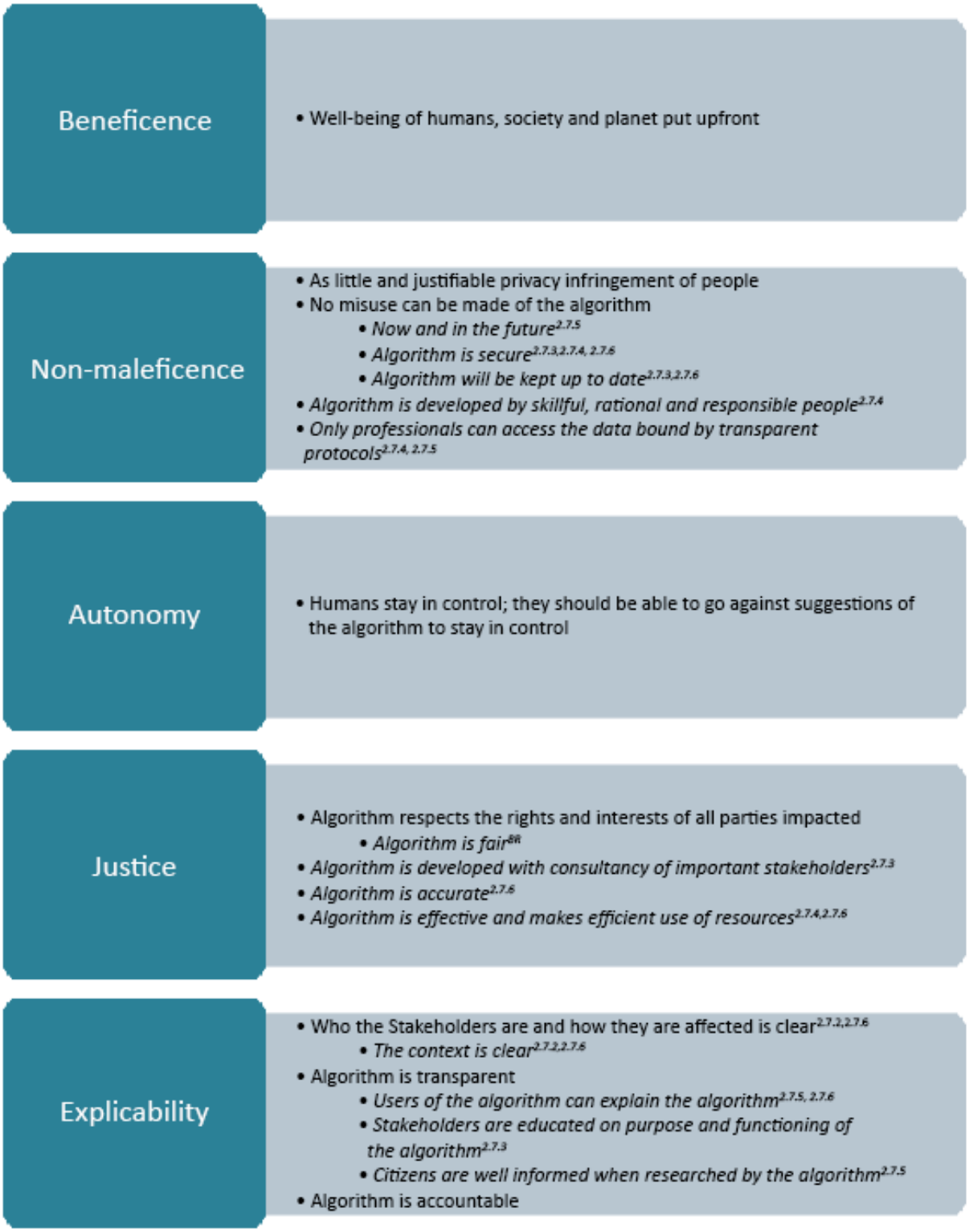


Figure 21 - Project framework

Please note that the starting question, 'Why do we use an algorithm?' is not included in this framework, since it is assumed this will be answered before using this framework. It is worth mentioning that during the last few weeks of this thesis, de Algemene Rekenkamer published a report on ethics and algorithms (Algemene Rekenkamer 2021). The framework published by them looks much alike the framework displayed in the project framework in Figure 21. It can be seen as proof that the topics of ethics and algorithms is a rather important one in the current moment. Differences lie in that the project framework also puts emphasis on the beneficence, rather than only the other four aspects. The framework of de Algemene Rekenkamer seems more focused on the non-maleficence and privacy side than on the beneficence side. Hence, there seems to be more focus on the 'do no harm' side than on the 'do good' side. Furthermore, the project framework seems to have more focus on the process surrounding the algorithm, like the development phase and communication to and education of the public. The framework of de Algemene Rekenkamer has more focus on the legal side, something that has been done in sub question 2 of the research, rather than in the ethical framework. So, in general the project framework, especially the extended version, seems more elaborated. It might be a good idea to merge these frameworks to some extent, since the project framework has a more theoretical approach and framework of de Algemene Rekenkamer a more practical approach. There should be looked per question of the framework of de Algemene Rekenkamer whether they add something to the project framework or whether they are merely a reformulation of the question. There should also be thought whether adding all legal aspects is something that is desirable. Currently, it is unsure how the framework of de Algemene Rekenkamer performs in the real world, that would be very interesting to compare too. This is beyond the scope of this thesis at this point in time, however it will be discussed further in 9.2 Discussion and future research directions.

6.3.1 Beneficence

This subsection focuses on the benefits of algorithms. According to Floridi et al. (2018) the principle of creating AI technology that is beneficial to humanity is expressed in different ways [in different frameworks], but it typically features at the top of each list of principles. It is, of course, not set in stone what 'do good' exactly means. There are many theories on what is good too, for example utilitarianism (best outcome for majority of affected citizens) or virtue ethics (what would a 'good' person do?) as can also be seen in DEDA. It is up to the organisation to choose an ethical theory to follow. This view should also be taken into account when using the rest of the framework.

Foundations for human beneficence can be found in human rights declarations and conventions. Since the interpretations may change over time and since laws are not always inherently good or may not inherently lead to human well-being, these declarations and conventions must be further evaluated and elaborated upon. This point will be further explained under explicability. What exactly should be of beneficence to humans must also be thought about. In case of social security assistance this leans towards the fairness of the social security system as a whole and to helping people getting back on their own feet (to work). It is important to look beyond the economic benefits, to subjective well-being (The

IEEE Global Initiative on Ethics of Autonomous and Intelligent System 2019). An example of this was given by the law enforcer: 'If someone is committing Marktplaats fraud, the common way of handling this is stopping the social security assistance and giving a fine. This does not necessarily help these people. It might be better for these customers to be helped in creating their own (KvK registered) business.'. Another part of this point is the planet's wellbeing. This is a bit less inherently logical, but it is important that she is taken into account. The organisation must make clear what the benefits are exactly and these benefits must fit the organisation and their goals.

6.3.2 Non-maleficence

This subsection is similar, but contrary to the previous one. The focus does not lie on the positive outcome, but on a non-negative outcome. A special focus point within this subsection is personal privacy. There is always a risk of invasion of privacy involved when using any kind of data, the question in this part is whether these risks are acceptable risks. According to Floridi et al, (2018) it is not entirely clear whether it is the people developing AI, or the technology itself, which should be encouraged not to do harm. In this chapter both are taken into account.

6.3.2.1 As little and justifiable privacy infringement of people

Thought must be put in which information will be used and why. Even though certain information may legally be used, does not mean it is ethically good to use everything that is available. All information that are used must add something to the decision-making process while keeping the ethical side in mind. This means that there must be no data included without a specific purpose or data that can excessively hurt people. There is then also the question when to add a new criterion. When the results go up by 0.5%? Or by 1%? There is a thin line between using more information to enhance the results and using as much information as possible. The data used should thus be proportional to the goal (in other words; the impact must be justifiable). A good question to ask here is whether the same goal can be achieved in a different way. It must thus be clear which information are used and why this information is used and the organisation must stand behind these decisions. If the algorithm is learning, often a training dataset is needed. The question arises here whether it is fair to the people included in the dataset that they are used to train the algorithm, while, most probably, the dataset used for training is an older one and thus outdated. This means that not all people included in this dataset are still customers. A fair solution to using a training dataset would be to create a fake training dataset so no real customers are included. This fictional dataset should be representative of the real world to properly train the algorithm (as also mentioned in multiple frameworks). Another option would be to anonymize the data. Either way, the most feasible option with the least privacy infringement must be chosen.

Furthermore, customers must be protected from being a false positive, meaning that the algorithm marked them as potential fraudster, while they are not. Of course, false positives will always occur, but the impact on someone when they are marked as potential fraudster is very high. The municipality may then use other methods, like house visits or conversations

with neighbours to see whether they indeed committed fraud. This may be humiliating for customers, so it is important that this, as far as possible, only happens to people who actually committed fraud. The citizen must thus be protected from being a false positive (too often). Furthermore, the exact impact being a false positive may have must be specified and documented. These are also discussed under Justice.

There is also difficulty in knowing what is little privacy infringement now and how this will hold up in the future. An example of this is anonymization: something might be considered anonymous now, but the question is whether it will still be anonymous in 20 year, or combined with other (open source) datasets or background knowledge for example. This is not something that can be found out now; we do not know what the future holds. The future perspective should be kept in mind when choosing which data to use, however. DEDA states that it might be good to ask this question ('can you imagine a future scenario in which the results of your project could be (mis)used for alternative purposes') to the project members.

Just as in the privacy part of this report, it is advised to conduct a DPIA to get a better understanding of the data and goals (and possibly hierarchy in goals). This also helps for the explicability part later on in this framework. As mentioned in DEDA, it is also advised to discuss all ideas (regarding this entire framework) with a privacy officer, if the organisation has one, or with an independent organisation.

6.3.2.2 No misuse can be made of the algorithm

Misuse can be performed in three different ways according to Floridi et al, (2018) namely in accidental misuse (overuse), deliberate misuse (misuse) and underuse. To prevent deliberate misuse, the security must be up to date. Accidental misuse can happen by users of the algorithm who have another goal for using the algorithm. Preventing underuse depends on getting the important stakeholders to see the importance of an algorithm.

To minimize misuse from the inside, clear protocols must be in place with rules for the algorithm. It must be clear what the algorithm is allowed to do and what it is not and how humans should interact with it. For example, it should be clear when humans are allowed to intervene or question the algorithm. This is closely related to the goal(s) and context of the algorithm. The protocols must therefore partly absorb risks and define boundaries.

6.3.2.3 Now and in the future[2.7.5]

As mentioned in the privacy part of non-maleficence above, misuse that can be done, changes over time. New technologies might be invented or security bugs may be found that were not known before. It is also possible that people look at the data who have background information which can compromise anonymity, for example. Therefore, here too, it is good to re-evaluate potential forms of misuse regularly to check whether something has changed since we cannot predict the future earlier than when it becomes the present. Citizens must be able to (rightfully) assume that thought is put into this.

6.3.2.4 Algorithm is secure[2.7.3,2.7.4, RW]

The security of the data must be ensured. Data that are used by the algorithm are often in plain text and almost always contain sensitive information. This makes it sensitive to hacking and exploitation, amongst others. Developers should be trying to detect as many ways of potential misuse and minimize those risks, according to IEEE. These risks do not only lie in the outside world, since users of the algorithm may also misuse them (accidentally or not). The users of the algorithm could for example be prevented from accidental misuse by the system creators by giving them only access to parts they actually need or by logging access.

6.3.2.5 Algorithm will be kept up to date[2.7.3, RW]

The algorithm must regularly be re-evaluated and updated to keep the algorithm reliable, consistent and within the set boundaries (ECP 2018). For this it is important to understand what 'good behaviour' of the algorithm entails to know how the algorithm should behave. It is also important to understand how exactly to check upon this. The algorithm must also not do any additional harm. Included in this is that the data must also be kept up to date. There is a moral and legal responsibility for keeping data up to date for both the customer and the municipality. The municipality can, however, remind the customer of their moral duty by sending emails or letters to remind them. If the data is not kept up to date, both the municipality and the citizen risk to evaluate based on incorrect data which may lead to incorrect results. The data must also be well organised and of good quality. The principle 'garbage in, garbage out' counts here. How the quality is checked and kept qualitatively good must be documented as well. Last, also the protocols surrounding all this must be kept up to date.

6.3.2.6 Algorithm is developed by skilful, rational and responsible people[2.7.4]

Professionals in data science are not always educated to also know much about securing these data. This is a risk that should be taken into consideration. Since the people who created the algorithm have quite some responsibility regarding the algorithm, these people should be skilful to ensure a well-functioning algorithm. They should be rational to make objectively good decisions. They should also be responsible, to be aware of the potential risks (including discrimination and hacks) to try and prevent them. The algorithm should also be tested extensively by these professionals, to ensure proper functioning. Development should be documented properly to make the process traceable and validated accordingly with, for example, test cases. What exactly is expected from these people may be specified in a code of conduct.

6.3.2.7 Only professionals can access the data bound by transparent protocols[2.7.4, 2.7.5]

Not everyone should be able to access the data, only professionals should be able to do so. IEEE mentions that access should be granted on a case-by-case base. With these types of algorithms, it is often the case that the data is encrypted or stored in multiple databases (for which a key is needed to couple these databases). The people who can access the data and especially de decrypted data or the keys must be professional. It is, for example, logical that there are less people who can access the full database compared to the people who can access suspects. There should not be too many people who can access these data. Finally, it

is good to have the protocols described in a transparent manner. This means, who can access the data and when is clearly defined, written down, and not kept a secret. This should be clear at all times. Another part to think about is how to monitor access and changes made by these professionals. The IEEE general principles also add to this that it might be a good idea to let a regulatory body oversee the entire process of using algorithms.

6.3.3 Autonomy

The careful reader will miss one point from the original framework in 3.2.1, namely that the stakeholders have the right to decide for themselves. In most other contexts this should be included, but with fraud detection, citizens cannot choose whether they may be investigated or not; that would forego the point of the investigation. Decision power is normally included here, and according to Floridi et al. (2018), people must be informed before being able to do so. Since there is no decision to be made by citizens, this information provision is included in explicability for this framework.

Algorithms should not have the autonomous power to hurt people. Humans must be able to always stay in control. The main point is that humans should have a form of choice, also when using an algorithm. This does not mean that algorithms should never hurt humans. Hurting people is needed to stop crimes. However, this should not be done by only the algorithm and not every decision of the algorithm should be taken for granted. Protocols should be in place for when humans should interact with the algorithm. Humans are rather bad in making good, rational decisions and a 'good' algorithms should be able to do better (Kimbrough, Wu, and Zhong 2002). This means that letting humans defer too much from the affirm could not be beneficial. However, a human is needed to perform the last check and to look at the grey areas. This also counts for citizens; they also must know how and when they can challenge a decision made by an algorithm.

6.3.4 Justice

The fourth subsection mentions that (the development of) algorithms should promote global justice. The foundation for this can be found in the laws and regulations, which should be followed. Again, these laws and regulations are not always ethically good. This will be elaborated upon under explicability.

6.3.4.1 Algorithm respects the rights and interests of all parties impacted

All parties must be reflected properly in the algorithm. It is good to realize that the rights and interests of all parties are different and these should be included. A data scientist, for example, would want to make the 'best' algorithm that works. Whereas a (fair) citizen would want social security assistance to be as fair as possible. Another example is a social security assistance customer, who wants as much privacy as possible maybe. These interests differ and may bite each other, but should all be considered.

6.3.4.2 Algorithm is fair

The algorithm should have no bias and no discrimination should be present; hence, algorithms should be fair. This is already mentioned in the GDPR, but ethically it is also very

important that the algorithm is fair because of its societal purpose. An example was given by the interview with the law enforcer. People from non-Dutch origin have a house abroad more often. This does not make ethnicity per se a good criterion to use, since it can easily lead to discrimination. The algorithm should thus be as objective as possible. This can, to an extent and as mentioned before, be tested with a training dataset or a fake dataset. The choice for a certain dataset and its considerations should be documented. Another example of unintentional discrimination was given by a man who wanted a credit card, but he could not get one because his in-living son was behind on payments, so his address was blacklisted. Floridi et al. (2018) mention that an algorithms can probably be used to eliminate past discrimination in the long run. An approach for predicting potential bias is given in DEDA, where the question is asked what the project members expect the outcome to be. Potential discrimination factors should be documented, according to the ministry of justice and security (Ministerie van Justitie en Veiligheid 2019). This is often the same bias as learned by the algorithm. If a learning algorithm is used it is also good to compare the hypotheses and the results so the project members can learn from this. It is important to realize that without the use of an algorithm discrimination can also happen unconsciously.

Often, for social security fraud, themes are used. In this, certain criteria are taken to check upon. Examples are to check on people who did not request special social assistance for over 3 years or people of whom (one of) the parents are unknown. This is on the edge of the no discrimination principle. It might be that certain groups of people do not know the entrances for asking special social security assistance, or that certain groups often have an unknown parent than others. Maybe with an algorithm, these themes are detected by the algorithm itself, which would create a fairer system, but it might still target some groups of people more often than others rightfully or not. In any case, there must be justifiable reasons that introduce this potential unfairness.

6.3.4.3 Algorithm is developed with consultancy of important stakeholders[2.7.3]

It is almost always a good idea to consult important stakeholders when developing, but especially here since the goals of the citizens, municipality and the developers are so far apart. The dilemma between using as much data as possible to get the best results and the least privacy infringement plays a large role here. Other stakeholders that might be good to include here are the client counsel and other (similar) municipalities and projects. This side of the development should again be documented properly.

6.3.4.4 Algorithm is accurate[RW]

The algorithm must deliver results that are as accurate as possible and thereby make correct judgements. This does not mean that an accuracy of less than 100% is necessarily bad. Full accuracy is quite possibly never reachable, since fraudsters are always a step ahead of those trying to catch them. Hence the resources to do this are often also behind. One should, however, always consider that being marked as a potential fraudster has an impact on people. This means that they will become a suspect and will be researched by, for example, house visits. Therefore, it is important to not have too many innocent people suspected (false positives) and not to have too many guilty people not suspected. People who are not clear-cut suspects must still be marked as suspects, because if a new way of committing

fraud is found, this must also be learned by the algorithm. So, it might be a good idea to have the accuracy as high as possible while also picking out a few random data subjects to further inspect. Keep in mind that this infringes their privacy, so the relation to the larger societal goal should always be weighed against it. The exact harm that the (non-)accuracy of the algorithm may cause must be defined. In the end municipalities are always responsible for the decisions made by the algorithm according to the ministry of justice and security (Ministerie van Justitie en Veiligheid 2019).

6.3.4.5 Algorithm is effective & efficient[RW]

Effective means that something must be successful in producing the desired result. In this case the goal is to detect as much, if not all, fraud (by an algorithm) for example. The first question should be whether an algorithm is the correct tool for achieving the goal or that there is a better way or whether it can be achieved by hiring more human capacity. It might be good to create some kind of benchmark for this, so comparison can be made with other organisations (or municipalities). According to IEEE, focus must be placed on how to interpret these metrics for effectiveness and efficiency. Metrics like these are not clear-cut. Of course, one might compare how much fraud is detected. This should also take into account the amount of data used, the privacy infringement, the investment that was needed and so on. There are no standards for this yet. Another thing to keep in mind is the effectiveness of the results of the algorithm. Suppose the algorithm was created to detect potential fraudsters and, after the first run, it returns 400+ potential fraudsters. There is no way to check all these people. A solution might be to check the capacity within the municipality. If that is, for example, to check 30 potential fraudsters per quartile of a year, then it might be good to check the top 30 potential fraudsters only, or the top 25 potential fraudsters and 5 random data subjects.

The algorithm should make efficient use of resources. This goes for the money being put into creating, implementing and using the AI, especially since this is public money, but also for the power (energy) that is being used while doing so. For example, it is ethically unwise to place the algorithm on a very old server or on a physical server with lots of unused space left. Also, the manpower should be taken into consideration. If more manpower is needed with an algorithm than without, it might be worth reconsidering. It also relates back to efficient use of data, hence using the minimal amounts of data to achieve a goal.

6.3.5 Explicability

A small group of people can create an algorithm which impacts the lives of many. This is why the subsection of explicability is so important. Floridi et al. (2018) mention that we need to know what the algorithm is and does before deciding upon whether it is beneficent and does no harm. To decide whether it is just, we need to know who is responsible. This added subsection therefore completes the jigsaw.

6.3.5.1 Who the Stakeholders are and how they are affected is clear[2.7.2,RW`]

It must be clear who the stakeholders are, what their position is and what their expectations and wishes are, according to the AIIA (ECP 2018). This can be done by performing a

stakeholder analysis. Included in here must also be the employees at the municipality. Direct stakeholders are often easily detectable. Indirect stakeholders (for example society as a whole, in this project) are less clear. They are, however, involved in this project; their money will be used in the most justifiable way. How the stakeholders are affected must be addressed on ethical, social, economic and legal aspects. It is good to keep in mind that the law is something that must be followed; however, it does not mean that this is all that should be done. If we look at the SyRI case for example, it is clear that even though the SUWI decree gave them a pass to use certain data, this did not 'feel' good for many people. So many people even felt like this that a site was created against it. This illustrates that the law is not always necessarily (morally) good. Of course, laws must be followed, but additional thought must be put in the consequences for stakeholders on multiple aspects without it being too narrow. This changes also over time, so this must be re-evaluated periodically.

6.3.5.2 The context is clear[2.7.2,RW]

The context must be clear to know the rules of the game. As mentioned by Wynsberghe et al. (2013), it changes a lot whether personal data are used for marketing purposes or for fulfilling a contract for example or whether they are used for medical (sensitive) data or for permits. If the context changes, all the checks, like the DPIA, ethical and legal checks, must be performed again to check its validity again. The protocols for the rules of the algorithm should be based on the context and goals.

6.3.5.3 Algorithm is transparent

According to IEEE, in transparent algorithms it should be discoverable how and why a system made a certain decision (i.e., why someone is considered a fraud risk). The ministry of justice and security mentions three reasons why transparency is important (Ministerie van Justitie en Veiligheid 2019). These are gaining trust, giving citizens the ability to check on the government and help them in this way, and it can lead to citizens complying better to the rules and regulations.

As mentioned earlier in this report, there are two ways of explainable AI: post-hoc explanation and transparency design. Post-hoc explanation seems most feasible and maybe also most desirable. Transparency design often gives up some quality, which contradicts with the goals of detecting as much fraud as possible. It should be kept in mind that humans are not good at making decisions (as illustrated before), hence there is often already a performance gain when using an algorithm, also with an algorithm that implemented transparency design. Next to this, full transparency may not desirable because it creates circumvention opportunities for fraudsters. On the ethical side, however, it is of course better to know everything that is happening (transparency design). The author does not believe full transparency is always beneficial however, mainly because of the circumvention opportunity. There should always be some extent of transparency, because otherwise it cannot be checked whether the algorithm makes the right decisions or whether it indeed does not include a bias. Therefore, especially the decisions must be explicable, but not necessarily the entire process. Even though the algorithm must not necessarily be fully transparent, the algorithm must be testable.

The letter of the ministry of justice and security seems to agree with this (Ministerie van Justitie en Veiligheid, page 7, 2019). The ministry mentions that the focus of the information provision should be on describing the goal of the algorithm, which criteria made the decision (and any decision rules) and the type of data that are used (quality and how they are combined). It must be kept in mind that training an algorithm is often not a transparent process, since we do not know what and how the algorithm learns exactly. Human learning is also a very complex process and humans can also make decisions based on gut feeling, so these problems already occur without the use of an AI. *How do we know why person A likes the colour blue, while person B likes the colour red?* Hence the aim should be to make the algorithm as transparent as human learning. The fact that algorithms become harder to fully understand may not be the reason for not trying to understand it nonetheless.

Another part of this transparency is not specifically focused on the algorithm itself, but also on the surrounding parts. All created protocols must be transparent as well. The protocols include who can access the data and when, but there must also be protocols for when something goes wrong (for example a data breach). It is also advisable to have exit and change protocols and protocols for what to do when the algorithm produces unexpected outcomes. In these protocols should also be included what data are used and what the general procedures are.

Most steps in these frameworks are there to make municipalities aware of what they are doing, why they do it and what impact it had, and to inform stakeholders, which should prevent outrage from the general public. It is still possible that, for example, citizens react very heavily on the algorithm and the municipality should think about that possibility and about what to do if that happens (and who is responsible for this).

6.3.5.4 Users of the algorithm can explain the algorithm[RW, 2.7.5]

To prevent misuse, the users (for example law enforcers) of the algorithm must know how it works and why the algorithm made a certain decision. This is an extension of ‘stakeholders are educated on purpose and functioning of algorithm' and ‘The algorithm is transparent’. The users should know more in-dept what the system does and does not. This is also important for checking whether the algorithm behaves properly.

Users of the algorithm must specifically be educated on how dependent and reliant they can be on the algorithm. Sometimes humans become too dependent on algorithms and things must be prevented. As well as becoming too reliant or too confident about these algorithms.

6.3.5.5 Stakeholders are educated on purpose and functioning of algorithm[2.7.3, 2.7.4]

According to Partnership on AI, the public must be educated and should have a say. When talking about social security fraud, the public is always a (indirect) stakeholder. On the other end of the spectrum is the fact that a social security algorithm cannot be completely transparent, since it would then make it rather easy for citizens to circumvent being caught for fraud. Hence a way in-between must be sought, where stakeholders are informed and educated to the maximum extent without being able to circumvent detection. Included in these information should be what is the goal of the algorithm, why a certain type of

algorithm is used, which data are used, how often the algorithm will check these data, what consequences may happen, who is responsible for the analysis and which quality checks happen (Ministerie van Justitie en Veiligheid 2019). If simple decision trees are used it must also be clear why a certain threshold was chosen. Someone should be responsible for communicating with and to the general public and other stakeholders. This should be done in a concise, understandable and easily accessible manner. According to the ministry of justice (Ministerie van Justitie en Veiligheid 2019) and security, the public must be informed via the municipality's website about:

1. The fact that the municipality performs data analyses,
2. Why they do this,
3. What consequences may occur for affected citizens,
4. Whether or not machine learning is used and an explanation of this,
5. What the legal basis is,
6. Which data sources are used,
7. Who is responsible for the analysis,
8. What the role of third parties is in the process,
9. What checks on quality are performed,
10. If there is a human intervention in the process and
11. Which assessment frameworks are present and how these are used.

Here we can take the analogous thought experiment. Suppose you are taking a shower and your Roomba would come in to vacuum the bathroom floor. Would you mind? Probably not. Suppose now this Roomba has a camera to more accurately see the walls and other bumps. Would you mind now? What if the camera would back-up to the cloud? Or let's go one step further. What if a blind and deaf man walks into your bathroom? Except for questioning how this happened, he cannot do much, since he received no information. But what if the man is not blind and deaf, and he walked into the bathroom to see you showering and to share this information, but when he runs outside to share this information he gets run over by a truck and dies, so he cannot share the information. Would it still be invasion of privacy? This shows there is a thin line between innocent algorithm and not-so-innocent algorithm depends on the information received, how this information is received and what the goal is and how important knowledge about the topic is.

Other important stakeholders are the employees at the municipality. If they do not see the importance of using an algorithm or if they do not understand the algorithm, it is difficult to get a project up and running. Therefore, the employees must be well-informed and their concerns must be addressed properly. This goes beyond the users of the algorithm and includes other employees as well. If they have any underlying feelings of uneasiness, it is also important to talk about this (DEDA).

6.3.5.6 Citizens are well informed when researched by the algorithm[2.7.5]

Citizens must especially know when they are marked as potential fraudster. They must also have the option to raise objections to the decisions made by the algorithm, according to DEDA. Citizens should at least be informed at the beginning of receiving assistance and when something changes. It would also be nice to remind them once in a while of what is

happening. This can be combined with a yearly request to keep their data up to date. Customers must also be informed when their data are at risk (even when this risk is very small or almost negligible) or when their data are looked at more extensively than the regular checks (so when they are a suspect).

6.3.5.7 Algorithm is accountable

It should be clear who made the algorithm and who is responsible for the way it works (Floridi et al. 2018). This includes the responsibility of choosing the data that are used, for how the algorithm works and for what to do in case something goes wrong, but it also includes the responsibility of creating protocols and other peripheral matters. Also included should be why the algorithm was created in the first place, why certain data were chosen and who can access which data (and why). It should even be known who is ultimately responsible for the entire project (see DEDA). This should be documented carefully.

The framework is created such that it is a checklist for organisations. Not all questions must be answered with a clear yes, but if they are answered with a no (or a 'yes, but') there should be thought about it. It is, for example, very hard if not impossible to make the algorithm completely unbiased, but there should at least be thought about how to prevent as much bias as possible. All points of the checklist should be considered and documented so it is clear later on as well. Hence, the framework should help organisations to think about the important ethical aspects of an algorithm or AI project too. This does not mean that this framework should not change in the future (when ethical viewpoints change) or when organisations believe certain point should be added. The conclusion whether the algorithm is considered ethical (after following the framework), how these decisions were made and how the decisions are justified must be included. There is no clear answer to whether something is ethically 'good', since it depends on the viewpoints and opinions of the people judging it and the organisation itself, and these may even change over time. Other people and organisations may have different opinions, but after the considerations of this framework and documenting those properly it becomes more substantiated. As mentioned throughout the framework as well, it is important to regularly re-evaluate the viewpoints to see whether they still match the organisation, society and other (implicit) rules. The full extended framework can be found in Appendix C.

6.4 Validation

The framework developed in the previous part will be validated to see how it holds up against practical cases. First, it will be validated with SyRI, because substantial information is present to answer the questions. After this, the framework will be analysed by applying it to the case of Gemeente Amersfoort. Not all questions in the framework may be answered because of a lack of inside knowledge by the author, especially in the case of SyRI. Due to time constrains the framework could only be validated with these two cases. These two applications are not enough to be statistically valid, but it does give an idea. The two applications are rather different from each other; hence some bias should be filtered out. How to validate this more (and better) will be discussed in future work. The full, filled-in, evaluation list can be found in Appendix D and Appendix E.

6.4.1 SyRI

First some methodical notes: during answering the questions of the framework, it became clear that some questions needed to be reformulated. There were some spelling errors and some open questions were changed into yes/no questions. For example, 'Who are the stakeholders of the algorithm?' was changed to 'Have you identified the stakeholders of the algorithm?'. This was changed while filling in the framework, because of three reasons. First, it did not change the framework, merely the wording. Second, it was otherwise not possible to draw conclusions. Third, it made the questions more clearly defined, which left less room for own interpretation and therefore should be more objective. One bigger change was made with the question 'Is it clear how they are affected?' (Explicability – Stakeholders). Earlier this included the ethical, social, economic and legal aspects as separated parts to answer. When filling in the framework, the author found it incredibly difficult to actually answer all these parts separately from each other and it was noted that some stakeholders are not affected on all aspects, which made it even more difficult. To make the framework more workable in practice, this has been changed to exclude the specific aspects.

One change should be made to the framework, that would change the framework and has thus not been changed during the evaluation. This is the question 'Is it clear who is responsible for the way the algorithm works, the data that are chosen, what happens when something goes wrong and creating protocols?' (Explicability – Accountability). This question should include the responsibility for the usage of the algorithm as well, since applying the algorithm is also a decision. This has been altered in the final version that can be seen in Appendix C, hence has been evaluated in its un-altered formulation. The fully filled in framework can be found in Appendix D, in Table 14, a summary of the results can be found of which a pie chart is displayed in Figure 22.

Table 14 - Summary of filled in ethical framework on SyRI

	☑	☐	☒	Not answerable (☐)
Beneficence	4	1	0	0
Non-maleficence	10	10	3	9
Autonomy	0	1	1	1
Justice	4	4	4	3
Explicability	11	5	9	11
Total	30	22	17	22

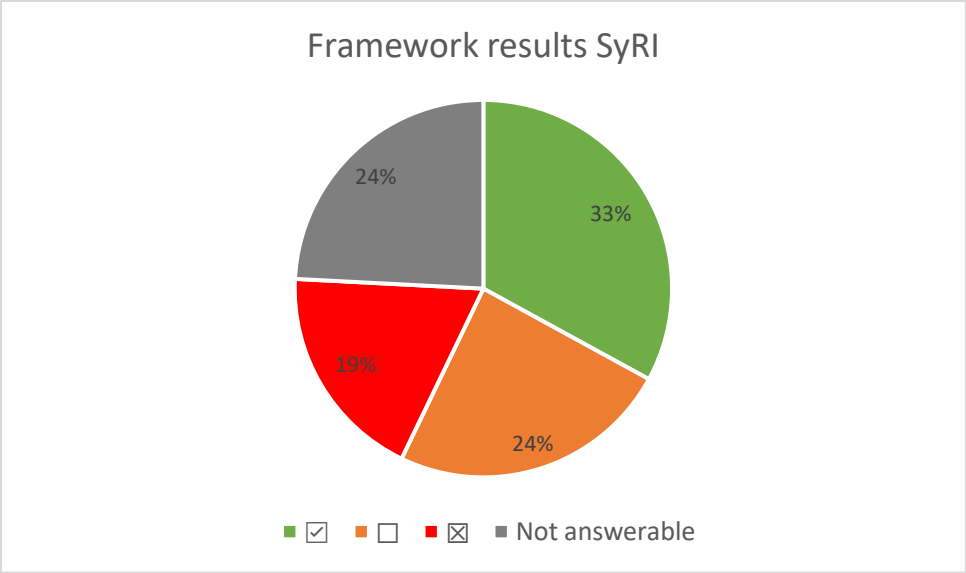


Figure 22 - Framework results SyRI

From the filled in framework it became clear why SyRI was created in the first place: the benefits are clearly present. To stop fraud from happening and to (re)gain trust from society in the system. A lot of money is lost due to fraud as well, so it is even a rather large purpose. Achieving this goal with the use of an algorithm also seems like the good way to go (see justice), because of the efficiency. In beneficence already, the first omissions appear, namely that the SyRI law was created with SyRI already in mind. Due to this, as can be seen in non-maleficence, practically all personal data that can be thought of is allowed to be used. It is unclear how much data fields add to the decision-making process, so it is unclear how 'bad' this is. Furthermore, it can be seen that the impact on someone when being falsely accused is huge. Good points here are that the algorithm has been subject to a DPIA and was reviewed by independent parties.

From autonomy, it became clear that transparency is a big issue with the SyRI algorithm. This is more thoroughly discussed in the Explicability part of the framework. A large issue that came forward was that there is practically no information to be found about the development process of the algorithm. It is thus unclear how it was created, but also whether the interests of other parties were taken into account or not. Another problem with SyRI, identified in justice, is that there was no proof that it would actually work. Two cases were made public, of which one had only false positives. There are, however, similar projects that have been proven to work.

Last, in explicability it is noticed there are a large number of questions that were answered negatively or could not be answered at all, due to a lack of information. The stakeholder part is not very representative, since it is unclear whether the people behind SyRI performed a similar analysis or not. What is noticeable here is that innocent citizens are also analysed by the algorithm, since it works on neighbourhood level. Interesting is that it is not desirable to have full transparency, because citizens might game the system. This, however, does not mean that no transparency is allowed. What was mainly marked as 'wrong' here, are the parts of educating the public. The public has only been limitedly informed and they are not

able to find more information online, for example. They would not know where to go to, in order to gain more information. People marked by the algorithm would also not know this and are not able to oppose the algorithm's decision, simply because they do not know about the decision. It is also unclear when exactly the algorithm is used, since the period it may be used after sending the informative letter is two years, which is rather long. It is, of course, a bit ironic that half of the questions that could not be answered were under explicability.

In short, the goal behind SyRI is clear and good. There is something to say for not having full transparency in this case. However, there are some parts of SyRI that were not handled so well. These were mainly the amount of data used, the accuracy of the algorithm, the impact on someone when they are a false positive, the transparency of the development phase and the education of the public. The entire framework should be re-evaluated periodically because of changing ethical norms.

6.4.2 Performance of the project framework on SyRI

The results of the previous part on SyRI will be compared to the court case about SyRI and the opinion on SyRI of the general media. Eventually ethics should resonate with the opinion of people and laws should reflect this opinion as well. The comparison was done to see whether the framework missed important parts and to verify that the framework is complete.

It should be noted that it is a bit different when filling in the framework for SyRI compared to when one would normally do this. Normally the framework would be filled in during or soon after the design or development phase. The author filled it in way after this period, even after the court decision to stop SyRI. What was also noticed is that a lot of information was unknown to the author. If the framework is filled in by the organisation itself, this should not be the case. The full comparison between the framework, the court case and the media is displayed in Table 15. Checks mean they have been discussed, while crosses mean that something was discussed, but this resulted in a different opinion or conclusion, tildes (~) mean no real conclusion was found and empty mean this was not discussed. The discrepancies are the most interesting for seeing whether the framework worked. Since it was unclear whether Bijvoorbaatverdacht.nl found the pseudonymization instead of anonymization actually is a problem, this was not put into Table 15.

Table 15 - Comparison of the framework of SyRI to the court case and the media

Topic	Framework	Court case	Media
Reasoning behind SyRI	✓	✓	✓
SUWI legislation mentioned as issue	✓		✓
A lot of personal data present	✓	~	✓
Impact on subjects	✓	✓	✓
Reviewed by independent parties	✓		✗
Unclear how SyRI was created	✓		✓
Lack of accuracy	✓		✓
Lack of transparency	✓	✓	✓
Lack of education to the public	✓	✓	✓
Risk cases are not informed	✓	✓	✓
A lot of unclarity is present	✓		✓
Results of SyRI are unclear/unverifiable	~	✗	✓
Bias present in SyRI	~	~	✓
Human interference should be present	✓	~	✓

First, the court case conclusion is briefly repeated here. From the 4.3.1 chapter of this report it became clear what the opportunities and issues were according to the court case. The reason behind SyRI was good. There was a lack of transparency, according to the court case[28], especially about which data are used and information provision when someone is considered a risk case. impact on someone when they are a risk case is huge. These did not weigh against each other. Important is that it seemed okay not to provide all rules regarding a decision of the algorithm to everyone. When someone is considered a risk case, they should know why the algorithm decided they are.

It is also important to compare the results with the opinion of the general media. Bijvoorbaatverdacht[18] is a good example of an opinion as mentioned. Six points are mentioned by them as issues. First, they mention that creating the law surrounding SyRI, while ignoring advice from important parties, was not good. Second, they mention it is weird that citizens cannot see their data. Third, the large amounts of data are mentioned by them as problematic. Tweakers highlighted this too[34]. Fourth, they mention that data used by SyRI

[34] https://tweakers.net/reviews/7452/2/moet-syri-worden-stopgezet-fraude-bestrijden-met-een-algoritme-inzet-en-kritiek.html

are not anonymized, but merely pseudonymized. It is unclear whether this is mentioned as a necessarily bad thing. Fifth, they mention that the results of SyRI are unclear and unverifiable. Sixth and last, they state that the economic benefit of the algorithm is relatively low. In another article[35], they mention that they think it is weird that the focus of SyRI was put on so-called problem neighbourhoods and the large impact being considered a risk case has on an individual. In here the lack of accuracy is also mentioned as an issue. The largest issue is that they feel that every citizen is already a suspect before they even did something wrong. Computable.nl[36] added an interesting highlight, which is that social benefit checks via an algorithm like SyRI indeed make the citizen suspicious in advance, however, speed checks on the highways and a similar algorithm of ABN Amro, a Dutch bank, do the same, without people complaining about it.

According to Bits of Freedom[4], who awarded SyRI the cynical Big Brother privacy award, the algorithm would create a stigma to the neighbourhoods it would be applied to. They also mentioned that citizens could not defend themselves. Amnesty[37] mentioned to have the most issues with the fact that citizens are kept in the dark and the lack of transparency of the algorithm. Interestingly, they mention that the algorithm cannot be checked, while the state mentioned during the court case that it can be. In the reactions on the earlier mentioned Tweakers article, people mainly want human interference with the algorithm, as in, the algorithm should not make decisions without a human checking these. Another part that is discussed, is about the lack of education of the public, but also of the governmental people who accepted the law for example. An interesting addition is done by computable.nl, who mention to find it beneficial to have the goal communicated clearly. As well as to have information on the development roadmap.

A few things stand out. First of all, much less has been discussed or concluded on during the court case. This makes sense, because they only ruled on a smaller part. The court can also not make assumptions, whereas the media can. The court also often give the benefit of the doubt, whereas this would not always be the case with the framework. Second, the main point of Bijvoorbaatverdacht.nl was that citizens would be suspect in advance. This has been placed under impact on subject, as that is what it is about. Third, the reactions on Tweakers.nl were very clear to want human interaction present. This has been included in both the framework and the court case, but it has not been judged on this by the court. Hence, in the framework this has also not been done to keep it as objective as possible and since not more information about this was present than what was said in court.

---

[35] https://bijvoorbaatverdacht.nl/gedachte-burgers-potentiele-fraudeurs-zit-systemen-als-syri-gebakken/

[36] https://www.computable.nl/artikel/opinie/computable-next/6829101/1509029/syri-ligt-nu-al-onder-vuur.html

[37] https://www.amnesty.nl/wordt-vervolgd/je-moet-soms-flink-tegen-de-wet-aan-duwen

There were four discrepancies. The first one being about the amounts of personal data present. The media and the framework found this to be a lot. The court also mentioned this briefly, however, they did not judge on this part. Still, the framework mentioned that it feels like a lot of data, and the media agree with this. The second difference is on whether independent parties reviewed the system. This was not a very black and white case in the framework, but it was concluded that it has been reviewed by multiple governmental people. The media seem to agree on this, however, they mentioned that many of these governmental people are not educated enough to make decisions on this. There is something to say for that, looking at the number of questions that were asked about SyRI. Nonetheless, there were some critical questions asked by governmental people. Competence of (certain people within) the government is often something of debate and it seems not to be different here. In the end, society chose the people representing them and the author believes this to be sufficient. Third, the media mentioned that the results of SyRI are unclear and unverifiable. The court case does not agree with this. Here it was mentioned that the algorithm is not more than a simple decision tree, and this should be transparent for the people working with it. The author believes the most objective way of looking at this is that the truth is probably somewhere in the middle. Fourth, the media often mention that SyRI is discriminating. It is difficult to say something about this since so much of SyRI is kept a secret. The court left this somewhere in the middle and the author chose the same approach with the framework, with a mention that it does feel weird that SyRI has only been applied to vulnerable neighbourhoods. This conclusion might have been different if the framework was filled in by a team member of SyRI, with more information. For example, the question 'Are there any potential discriminating factors in the algorithm?' is one that should be filled in then.

For this specific case, the framework was found to work, since it covered all the flaws found by the court case and the media. These were the amount of data used, the accuracy of the algorithm, the impact on someone when they are a false positive, the transparency of the development phase and the education of the public. Comparing with the court case is a bit biased, since much of the information used to validate was received from the court case. Hence, whether the results of the framework correspond with the media might be more interesting. All focus points of the framework were found in the media as concerns as well. The other way around this was not the case, often because the assumptions were made by the media and not by the author. If more information would have been present, this could have been avoided. This was also one of the main points that came out of filling in the framework.

This first validation did filter out some impracticalities of the framework, as mentioned in the first paragraph. If the framework would have been filled in when starting the development of SyRI, this could have pointed out the ethical flaws with time to resolve these as well. Of course, after one validation to a real-world example, these conclusions can only be made carefully.

6.4.3 Amersfoort

During the validation of Amersfoort, it was found that one additional change should be made to the framework. Namely the question 'Is the algorithm also privacy friendly for people with background knowledge or when combined with other data sources?' (Non-Maleficence – Privacy) can almost never be answered with a yes, especially not for these fraud detection algorithms, since the data is never fully anonymous. Even if the data were to be 'fully anonymous' it would still be possible to recognize exceptional situations, such as people with many children. Adding the word 'reasonably' in this question would solve this. This would make room for ensuring anonymity by other means, like a processing agreement, an NDA or another type of contract. This would also make this specific question in both the SyRI validation and the Gemeente Amersfoort validation green instead of orange. This will be altered in the final version of the framework that can be found in Appendix C. The fully filled in framework can be found in Appendix E. in Table 16, a summary of the results can be found of which a pie chart is displayed in Figure 23.

Table 16 - Summary of filled in ethical framework on Gemeente Amersfoort

	☑	☐	☒	Not answerable (☐)
Beneficence	5	0	0	0
Non-maleficence	22	8	1	1
Autonomy	2	1	0	0
Justice	11	2	2	0
Explicability	23	8	3	2
Total	63	19	6	3

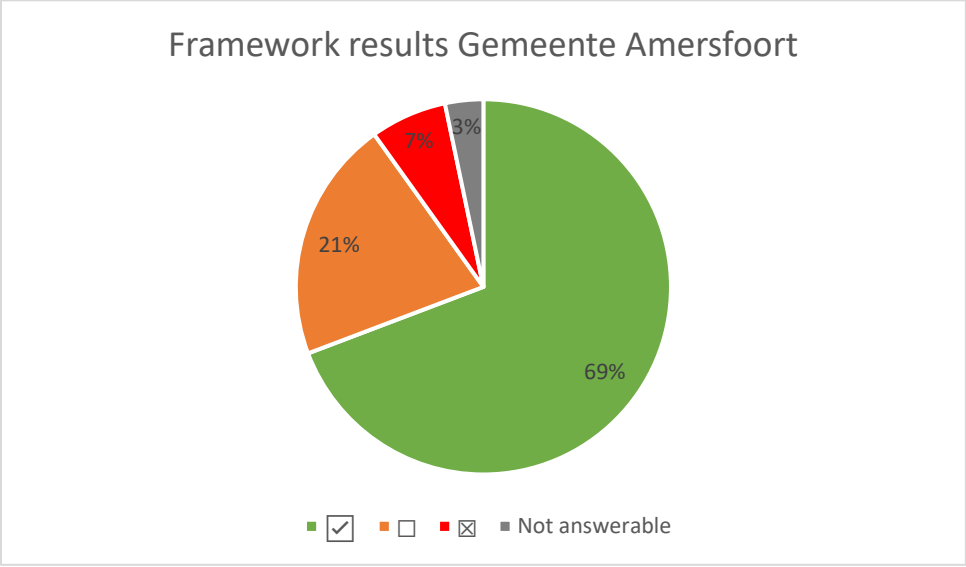


Figure 23 - Framework results Gemeente Amersfoort

Before starting the description of the results of Gemeente Amersfoort, it is important that these results are not precisely comparable to SyRI. This is mainly because 19 more questions were answerable. It is not reasonable to assume that these 19 questions would all have been answered positively for SyRI for example, so that is important to keep in mind. Another difference is that the SyRI algorithm was already is use, whereas the algorithm from Gemeente Amersfoort is still in the development phase. This means that some aspects simply have not been arranged yet, because they were planned for a later stage. This means that questions are marked red or orange in the framework and when they would be re-evaluated later on, these may turn green instead.

Just like with SyRI, it is clear why Gemeente Amersfoort would choose for a fraud detection algorithm: to stop the occurrence of fraud within the social security assistance context. This, in turn, to increase trust in the government and the tax system as a whole. The mindset of Gemeente Amersfoort is also very special, because they feel that the earlier fraud is detected, the earlier people can be helped. They also want to rule out the human differences between law enforcers.

From non-maleficence can be seen that there should be awareness of the fact that the algorithm needs training. This is not necessarily something bad, but it means extra data is needed, which should be thought about. There are a few factors in the algorithm that may be considered sensitive information, such as gender and special notices. These were also identified by the team. It still has to be evaluated whether these will stay in the algorithm, and it should then be really thought on whether these criteria add something to the decision-making process. With that information in mind, some questions in the framework must be re-evaluated in a later stage, to check upon the ethics on these points again. These are located in Non-maleficence and mainly in Justice. It must not be forgotten that the impact on someone who is considered a risk case is large. This differs from SyRI in that the customers are always first investigated behind the desk, before they may be invited for a conversation to research the suspicion. The research will only continue if there is suspicion by the law enforcers. The impact is not necessarily larger than without an algorithm, but this is always something to keep in mind. This will also be further discussed in Chapter 9. Next to these, some protocols must still be altered. This is planned to be done in a later stage; hence this should also be re-evaluated. This is also mentioned in Autonomy and Explicability. Last for non-maleficence is about the risk of this algorithm. This risk is present because the algorithm has not been tested, since it is a relatively new algorithm. However, without real-world cases applying the algorithm, there will of course never be any innovation. Another thing to consider is mentioned in Justice, which is that there is always some inaccuracy present with the current algorithm. This is not necessarily something bad; it keeps the law enforcer objective, since they know that of the 12 risk cases the algorithm mentioned, 2 are random and therefore likely to be non-fraud. They do not know which 10 cases were indicated by the algorithm nor which 2 cases were provided at random. The main reason is that they cannot completely rely on the algorithm in this way.

Last in Explicability, there are also some points to keep in mind. These are that the context could change. This might not be something bad, but one should be aware of this. Second, customers do not automatically get a notification why they are considered a risk case or what their fraud score is. They can probably request this, but it might be good to research how and to what extent this may be done. Law enforcers also do not know why someone was considered a risk case by the algorithm to keep them as objective as possible. Next to this, there are currently no protocols in case of an extreme public outrage. These protocols would probably be useful, since there is quite some resistance to these types of algorithm after what happened with SyRI. It is good to know what should happen in this case, by whom and in what manner. Another point that stood out from the framework is that the citizens of Gemeente Amersfoort currently do not know that an algorithm is being developed. They will get to know this after the council has been informed. There is something to say for this, namely that it is not sure yet whether the algorithm will actually be used. There is also something to say against this, namely that citizens have the right to know, to some extent, what occupies their municipality. Gemeente Amersfoort should make a clear decision on this and these questions should be re-evaluated after this decision or when the algorithm is being further developed. Related to this is the information to the public. Currently it cannot be said whether this is done in a concise manner, since it has not been created yet, so this should be re-evaluated later on. It was mentioned by Gemeente Amersfoort that they want to publish the technical documentation and that they want to include extra information with this. From the framework the recommendation is that this documentation should include why data analysis is performed, what the consequences are for citizens and the legal basis and an extension of the explanation of the quality checks. Last, customers have the duty to provide information to the municipality. This might for example be when they are moving in with a partner or when they get a new job. Customers are not currently reminded that their data is checked (later on maybe by an algorithm). There is some legal basis not to not inform customers of their data being checked, the question is whether the ethical basis is there too. Gemeente Amersfoort should think about whether this would be desirable to do and the accompanied question should then be re-evaluated.

In summary, the goal behind the algorithm of Gemeente Amersfoort is also clear and good. Many parts are arranged very well. There are some parts to be aware of for Gemeente Amersfoort however. These are the sensitive data that are included in the algorithm, the protocols that need to be updated or created, that citizens do not know about the algorithm yet, that the communication to the public must still be created in a good manner, that customers do not automatically get to know why they are considered a risk case and that it might be good to remind customers once in a while that the algorithm is used. There are also some parts that should be kept in mind, which are that the algorithm needs training, the impact on customers, the risk that is present with using a brand-new algorithm, the inaccuracy that is deliberately present and the fact that the context may change. As mentioned earlier, not every part of the framework must be green, but all parts should at least be thought about. The entire framework should be re-evaluated periodically because of changing ethical norms.

6.4.4 Performance of the project framework on Gemeente Amersfoort

It is, of course, a bit more difficult to analyse the performance of the framework on Gemeente Amersfoort, because no court case has found place and less media attention is present. The media attention that is present, is based on the algorithm of Gemeente Nissewaard mainly. This algorithm is also created by Totta Data Lab and is rather similar, however, different data fields are used for example. Therefore, there will be looked at what the media said about the algorithm of Gemeente Nissewaard.

The full comparison of all points mentioned in the conclusions of the framework and found in the media or during the interviews can be found in Table 17. Checks mean they have been discussed, while crosses mean that something was discussed, but this resulted in a different opinion or conclusion, tildes (~) mean no real conclusion was found and empty mean this was not discussed. The discrepancies are the most interesting to see whether the framework worked.

Table 17 - Comparison of the framework on Gemeente Amersfoort to the media and the interviews

Topic	Framework	Media	Interviews
Reasoning behind using a fraud detection algorithm	✓		✓
Some sensitive data present	✓		✓
Protocols need to be updated or created	✓		
(Lack of) Communication with the public	✓	✓	✓
Customers do not know automatically why they are a risk case	✓	✓	✓
No periodic reminder present	✓		
Algorithm needs training	✓		✓
Impact on customer	✓	~	
Risk of being an early adapter of such an algorithm	✓		✓
Inaccuracy present in the algorithm	✓		
Context may change	✓		✓
Users are able to explain the algorithm	✓	✗	
Human interference must be present	✓		✓
Which data are used and why are clear	✓	✓	✓
Algorithm should not make decisions for us	✓		✓
It is discoverable why the algorithm made a certain decision	✓		✓

FNV is the main adversary of the algorithm of Totta Data Lab at Gemeente Nissewaard. It is often mentioned in the media that there seems to be a lot of confusion. This because it was mentioned that 'a person with three cars who receives social security assistance is weird', while it should not be possible for Totta Data Lab to see this information[38]. There was also some confusion about how much fraud has been detected by the algorithm[39]. FNV has also mentioned that they believe that every customer should be informed what their scores are, not only those who are considered a risk case[40]. Mentioned by Volkspartij Gemeente Nissewaard (a local political party) is that not much has been communicated to the people [41]. They also mention that the technical documentation is very difficult to read. Platform Burgerrechten adds to this that they believe the algorithm to be 'random' and they believe the municipality should be able to explain why the risk cases are a risk[42].

During the interviews performed in 3.2, also some concerns were identified. Due to the lack of articles found in the media and the relevance of these concerns, these will be added to this chapter as well. There were also two interviews held that were not included in this report before, but now provide useful insights. These were with the information provision advisor and with the data scientist of Gemeente Amersfoort, both held on the 18th of March 2020. It is interesting to take their concerns into account here too, since they provide a different angle. The information provision advisor mentioned that they saw drawbacks of the algorithm in that the context of the algorithm may change and that the municipality wouldn't use such an algorithm, because it is new and might be scary. The data scientist mentioned three points. These are that the algorithm should not make decisions for us, it may be difficult to make the algorithm explainable and that these algorithms are new and still need to be proven. All concerns can be found in Table 17, the explanation of the concerns found during the interviews can be found in the interviews in 3.2 or above.

Here too, a few things stand out. There is much less information found in the media than during the analysis of SyRI. This makes sense, because there was more outrage at SyRI than there is at the algorithm created by Totta Data Lab. Next to that there are many points discussed in the framework which are not mentioned in the media or during the interviews. This makes sense in this case, since also inside information was present to the author, in contrast to SyRI. Not all this information is present to the general public (yet). Part of the concerns mentioned by the media or in the framework are combined under communication

---

[38] https://www.ad.nl/voorne-putten/fnv-op-ramkoers-met-Gemeente Nissewaard-over-fraude-opsporing-rechtszaak-dreigt~a4bd9b0c/

[39] https://www.ad.nl/voorne-putten/Totta Data Lab-data-lab-blijft-bijstandsfraude-opsporen-ondanks-ondoorzichtige-controles~a87b4743/

[40] https://www.fnv.nl/nieuwsbericht/sectornieuws/uitkeringsgerechtigden/2020/06/fnv-sceptisch-Gemeente Nissewaard-belooft-burger-transparan

[41] https://www.binnenlandsbestuur.nl/sociaal/nieuws/rookgordijn-rondom-fraudesysteem-Gemeente Nissewaard.13026523.lynkx

[42] https://platformburgerrechten.nl/2020/09/22/Gemeente Nissewaard-weigert-alsnog-beloftes-transparantie-na-te-komen/?s=Gemeente Nissewaard

to the public. These are for example that citizens do not know about the algorithm and the communication to the public must still be created, both from the framework, and the confusion mentioned and the unreadability of the technical documentation in the media.

Points not discussed by the media will not be focused on, because, again, it does make sense that the framework would be stricter and/or have more inside knowledge. It is, for example, logical that the media does not know whether protocols are up to date or not. Next to that, there are two discrepancies. First, the impact on the customer was not mentioned explicitly by the media, but it was implied sometimes[43]. The second discrepancy is that users are able to explain the algorithm. This entails that the decisions of the algorithm must also be clear. According to the media, in a reaction on Gemeente Nissewaard, this was not the case. It is unclear where this information comes from and whether this is based on facts or suspicions. The framework mentioned that users are able to explain the algorithm, also because Totta Data Lab worked on explainable AI within the algorithm. Last a lack of communication was mentioned in the media, but this was mentioned about Gemeente Nissewaard, not about Gemeente Amersfoort. It does still highlight that this is an important aspect of introducing this algorithm, as was also mentioned during the interviews and was found from the framework.

Another thing that is worth mentioning is that, while gathering the information needed to fill in the framework, it was sometimes mentioned that there had not been thought about, but that it might be something that should be added. So the framework worked well as a checklist of what aspects to cover. An example being a protocol for when public outrage would occur. Next to that, the author of this report also discovered new parts about the algorithm of Gemeente Amersfoort. For example, three more stakeholders were uncovered.

It is a bit harder to say whether the framework worked for this case, compared to the SyRI case, because there is less information from the media and there was no court case. Whereas the framework in SyRI had found fewer points than were mentioned in the media, mainly because some parts in the media were assumed, this was not the case for Gemeente Amersfoort. Here more information was present, which made the framework also identify some inside recommendations the public would not know about. There was also less public outrage, which led to less extreme news articles. It is important that what is in the media is reflected in the framework, but not that many articles were found and those that were discovered were written about Gemeente Nissewaard. This is primarily because it has not been communicated to the public about the algorithm of Gemeente Amersfoort yet. When this happens, the framework should be re-evaluated. Hence, the interviews are just as important. All what was found during the interviews was also concluded from the framework. The author of this report also identified some new parts about the project and Gemeente Amersfoort did mention there were parts that sounded interesting that had not been thought of before. Hence, framework can help to uncover aspects that need attention. Some issues identified by the framework were already planned to be fixed in a later stage. If

[43] https://bijvoorbaatverdacht.nl/Gemeente Nissewaard-belooft-burger-transparantie-over-fraudescore/

this framework were to be re-evaluated by then, there should be a solid ethical evaluation to work upon. All in all, the framework also seemed to work for this specific case and it provided Gemeente Amersfoort with some recommendations and points to keep in mind.

6.5 Conclusion

The framework was validated using two real-world cases and it turned out to be a well-working tool in both cases. Since it was validated using only two cases, more validation is needed to ensure its robustness even more. The framework needs to be applied a couple more times to be able to claim this with certainty. What is known, however, is that the framework did seem to work for the two cases of SyRI and Gemeente Amersfoort. These cases have a few things in common. First, they are of high societal importance. Fraud is an important phenomenon and large sums of money are involved. Hence the reason why the solutions were or are created are more than clear. Second, in both cases the data subjects, citizens or customers, do not have a choice to participate. They cannot say that they do not want their data to be used. This makes that their autonomy is limited. Third, both applications are something new, which makes it sometimes scary to try since the long-term effects are not clear yet. What can be seen from the previous results is that it does seem useful to fill in the framework during the development or design phase already by someone who is in some way involved with the algorithm and has inside knowledge about it. This mainly because then the results of the framework can still be used to improve the algorithm and its surroundings and because in this way all issues can be discovered. For example, in the SyRI case it is hard to believe that all issues were identified, because so much was unknown.

The framework is currently designed to be applied to algorithms with limited to no autonomy for data subject. If the framework were to be applied to a case where customers can choose whether they want to participate or not, the autonomy part in the framework should be adjusted to include ‘that stakeholders have the right to decide for themselves’. This was left out for these specific cases, but this should then be re-evaluated. Furthermore, the framework should be redesigned to be applied to something different than algorithms. This could, for example, be other technical inventions like apps. This is not the easiest, since the background for the framework was based on AI specifically. If this is desirable, there should be new research performed on this. This will also be discussed in Chapter 9.

For now, what can be concluded is that the framework did seem to work for the two specific cases, with the framework being especially useful when filled in during the early stages of a project. Interesting was mainly that the framework uncovered new aspects that needed (more) attention, that were not discovered before. This proved the framework to be a solid checklist of what aspects to cover. With the filled in framework, it also became very clear what did go well already, from an ethical point of view. What the validated cases have in common are the societal importance, the lack of autonomy for data subjects and that they are cutting-edge technologies in this context. If it would be desirable to apply the algorithm to algorithms in a different context where data subject do have autonomy, some adjustments should be made. Next to that, if the algorithm would be applied to a non-AI

context, the entire framework should be re-evaluated, including the literature research. All in all, the framework seems like a promising way of evaluating the ethical side of AI projects.

The ethical considerations that should be taken into account when using a fraud detection algorithm can be found in short in Figure 21 and extended in Appendix C.

Chapter 7

7. Effectiveness, Accuracy, Legitimacy & Justness

In this chapter the effectiveness, accuracy, legitimacy and justness of the conventional attributes used in fraud-detection algorithms will be looked into. The implications of these answers will be focused on as well.

7.1 Method

This chapter looks into whether the conventional assessment criteria of the algorithm, namely legitimate, justified, effective and accurate, and what impact that has on the designing phase of the algorithm. For the legitimacy part, the results of Chapter 4 are applied to the case of Amersfoort. For the justified part, the results of sub Chapter 6 are applied to the case of Amersfoort. To answer the part about effectiveness and accuracy, attribute selection is performed to the RF algorithm developed earlier. From here can be seen whether there are attributes that do not have real predictive power. To see whether extra attributes could be added, the interviews done in sub Chapter 5 are re-evaluated. Last, the implications of the results of these questions are highlighted.

7.2 Effectiveness

To answer the part on effectiveness, it must be clear whether all attributes add something to the decision-making process. Hence, this part will look into which attributes can be removed from the dataset without losing too much of the accuracy. Next to that, it also makes that less data are needed, which helps with the privacy aspect, as well as with calculation speed. Note that it is important that not too many attributes must be removed solely based on the training data, to prevent overfitting. Overfitting happens when the model is too well trained on a certain dataset, such that it impacts the performance when applied to a new dataset. This can be seen when the accuracy of the training dataset (on which the RF will be trained) would be much higher than the accuracy of the testing dataset (on which the RF will be tested). An example of removing too many attributes and introducing the possibility of overfitting, would be to remove fields with relatively new data, such as data on the Wet maatschappelijke ondersteuning (WMO). the After each step, the accuracy will be checked briefly; it should stay about the same as it was. After the final selection, the algorithm will be tested more extensively. This will only be done after the final step because of time and processing power.

In the earlier part on creating an algorithm (5.3 and 5.3.6), a look was already taken at what tables and attributes were most important. This part investigates which are least important. Some of the attributes that were used for the RF algorithm had very low predictive values because they were simply not present in the dataset used for training the algorithm. This is because some of the types of, for example, activities were only present for less than 10 cases in the entire dataset (15180 cases) and even less cases in the training dataset (1093 cases). It is logical that these attributes do not add much to the final prediction. In first instance, there were 24 attributes that had a predictive value of 0 in the RF algorithm. These were specific categories in the tables dossiers, steps, participation process, activities, contacts, debtors, special situation and RPO. From this, it was noticed that the attributes that did not add much to the decision-making process were attributes that were indeed barely present. Even when these, barely present, attributes all turned out to be fraud cases, this does not say much, since it could still just be sheer chance. Hence, it had been decided to remove or recode all attributes with less occurrences than the threshold. This threshold has been set to 1% of the training dataset, which consists of 766. Hence the threshold is 7 times or less.

It may be that some of these attributes were only limitedly present because of a newly introduced law, for example. This means that there will be more cases in the future. It might thus be good to re-check whether the threshold will be reached in the future.

If the column was numeric, for example count of requests, and had less than or equal to 7 occurrences where the count was non-zero, the entire column was removed. If the column was categorical, and one of the categories occurred less than or equal to 7 times, the category was recoded to the category 'unavailable'. The removed/recoded columns/categories and categories were present in the following tables: dossiers, contact moments, steps, participation process, activities, debtors, RPO, special situation and requests. In total, 12 columns were removed or recoded. The table from which the most categories were recoded was activities. Running the RF algorithm again with the altered dataset gives about the same accuracy. Note that Totta Data Lab removed categorical data if a category was present for less than 5% of the customers, so that is more than was done here.

Next to this, it is also important to look at the correlation of attributes. There are algorithms that perform better if these correlated attributes are being removed. Next to that, the attributes do not add much if they correspond too much with another attribute, since they add no new information. Removing them, again, helps with privacy and with calculation speed. Caret, the R library used in sub question 3, also has correlation calculation built in. For all attributes, their correlation with all other attributes was calculated. Then, the output was given when an attribute had a correlation higher than 75%, which often is the cut-off from where one would want to remove an attribute. Caret then looks at the attribute with the highest average correlation, and outputs this. There were 7 attributes found which all have a correlation of at least 75% with another attribute and had the highest average correlation. These were the *start of the latest activity*, *maximum age of children*, *number of requests*, *contact handled*, *number reply*, *code action*, and *code of latest special situation*. These were removed and the algorithm was ran again. This resulted in an accuracy that was about similar to the accuracy with these attributes included (69.72%).

To make sure the findings in accuracy are not simply a matter of luck, the algorithm has been run three times for the full data set and three times for the altered dataset, as described above. It was found that the full dataset had an average accuracy of 69.42%. The altered dataset (with attribute selection and correlation analysis) had an average accuracy of 70.13%. The most important point of the attribute selection was for the accuracy to not go down. In this case the accuracy even was raised by about 0.7%. This might be because the algorithm makes trees with attributes that have more predictive power; hence the trees might be more accurate which leads to the algorithm as a whole being more accurate. Furthermore, it was noticeable that the time it took for the algorithm to be trained was, on average, 35 minutes for the RF model with the full dataset and 20 minutes for the RF model with the altered dataset, which was 43% less time. This thus makes it more efficient, because there are, again, less attributes with low predictive power.

There are many ways to do attribute selection and the number of attributes can still be decreased drastically. Some people mention to only use the top 20 most important

variables, as seen by RF. Others call that overfitting. When testing this top 20 method, an accuracy of 66.97% had been found, which was a rather significant drop in accuracy. As mentioned before, the algorithm and dataset can be tweaked almost infinitely, but by this attribute selection it became clear that there are indeed some variables that can be removed. An addition to this could be to try the 'hold one out' method, which looks at whether the accuracy drops or not by leaving out one of the attributes at a time. Depending on this accuracy drop, the attribute could be left out or not. This is something that could be valuable with more time and processing power available. Furthermore, there are also categories that do not add much, so this hold one out technique could also be tested with categories. These changes are displayed in Appendix G. It can be said that this part indicates that a good look should be taken into whether all attributes add something to the algorithm.

So, the top 20 most important attributes for RF were already mentioned in 5.3.3 and the most important tables were discussed in 5.3.6. In this part it has been looked at the least important attributes and tables to simplify the algorithm. Five things stood out. First, especially the activities attribute had very many categories. Logically, these cannot add much to the decision-making process, since some of these categories only occur a few times. Second, when removing all features with less than or equal to 7 occurrences, the tables requests, activities, special situation, contact moments, debtors, dossiers, participation process, RPO and step were affected. Noticeable was that most affected attributes within these tables were categorical attributes. Third, when the correlation between attributes was examined, the tables special situation, requests, activities, contact moments and children were affected. Of these, the fields date handling, code action and number reply in contact moments and special situation code in special situation were removed. Fourth, one could consider removing more attributes, however, this introduces the possibility of overfitting, which should be avoided. Fifth, the accuracy did not go down and, in fact, even was raised after the attribute selection process.

7.3 Accuracy

This part looks back at the interviews that were held with the law enforcers to find out whether there are external data sources that add something to the decision-making process. Whether these should or should not be added to the attributes used by the algorithm will be discussed here as well.

The accuracy talked about in the previous part was already rather high. The exact accuracy of Totta Data Lab was not given by them. However, they do mention that, for the Gemeente Nissewaard algorithm, of the top ten predicted fraud cases by the algorithm, 50-60% indeed frauded. The accuracy found during this research was a bit higher, around 70%. If it is only looked at the fraud predictions, the accuracy is around 64% (18 false positives and 32 true positives). Although, the found accuracy is already quite high, it still makes sense to attempt raising accuracy by adding more attributes.

During the first interviews, held in March 2020, a question was asked on which data, apart from the data in the database, are normally requested in a fraud research. Mentioned here were data from Marktplaats, social media (especially Facebook and Instagram), Kadaster

(ownership and use of real estate), Belastingdienst (taxes), RDW (vehicle), bank statements, GBA (municipal taxes), customer managers and participation process mentor. The latter two will not be considered, since these cannot be automatically added. During the second round of interviews, it was found that law enforcers also used travel data and water usage in addition to the earlier mentioned data.

Apparently, the law enforcers use these data to complete their view of the customer, so there must be predictive value in these data. Some fraud can be easily detected by one of the earlier mentioned data sources. For example, cohabitation fraud can often be discovered by social media and water usage. Illegal work fraud can be detected by, amongst others, travel data and bank statements. However, as was seen earlier in the legal part of this thesis, combining these data sources is often not allowed. For SyRI, a special law was created to do so, but this law is not present for this case and is currently no longer there at all, because of the lawsuit. So, these sources can be requested manually if there is suspicion about a customer, but these cannot be automatically processed or coupled, which would be the goal of such an algorithm. It is also described in the GDPR, as was also mentioned in sub question 2, that there needs to be data minimization and purpose limitation. Both of these points will not be met when adding these data sources; people who give their data to, for example, the Belastingdienst do not expect their data to be used in a regular fraud check from the municipality. Even if this was allowed, question arises whether one should or even would at this point in time, because of all the controversy with algorithms in general currently in The Netherlands. Adding these sources might raise the accuracy, because the law enforcers use the sources as well (hence they must add something). If the time and place is right, this could be researched.

Suppose these legal concerns were not present, it might be an idea to add specific data to certain suspicions. This would, for example, create a request per fraud suspicion case to gather the data from the relevant sources. For example, if someone is suspected of illegal work, all relevant data would be requested and analysed by another algorithm. This would create a new algorithm which has a different goal and purpose. Questions arise whether the municipality must do this, or whether this should be arranged by the owners of the data sources or by the government for example. The potential, however, should not be dismissed.

Social media data (including Marktplaats data) are a bit different. Mainly because people post these data to be seen. For example, on Twitter, tweets are even emphasized with a hashtag to reach a broader audience. On Marktplaats people categorize their items to sell them faster. van Wynsberghe et al. (2013) proposed ethical guidelines for early assessment of ethical privacy concerns for researches dealing with social media data. These guidelines were already discussed in 2.7.2 and the exact model can be found in Figure 6. These guidelines can be applied to this case.

1. The stakeholders are clear and discussed in 2.5 Stakeholders. Indirect stakeholders are discussed here as well.
2. The contexts of the data collection are Facebook, Instagram and Marktplaats.
   Facebook is a rather private platform. Users must add other users as friends in order for them to have a connection. Furthermore, users have privacy controls and users do not

typically make use of hashtags. Facebook also indicated that their data is rather private by only selling it in aggregates. Facebook also has the options to make certain (sub)parts of one's profile private, as well as the entire profile. Hence, Facebook data is rather private.
Marktplaats data, on the other hand, is public by default. There is no way to make posts private either. Hence, users know their posts can be seen by anyone and everyone. Marktplaats is used by users who often use a username which differs from their own name. That makes it possibly difficult to recognize someone, however Henry Been created a model for identity resolution for Twitter, which seems promising to be used here as well (van Wynsberghe, Been, and van Keulen 2013). Marktplaats is already scraped by the UWV to collect data (Olsthoorn 2016). Hence, these data can be considered public.
Instagram is somewhere in the middle privacy-wise. Users can make their entire profile private (not subparts), thereby preventing others from seeing the content they post. However, Instagram has hashtags which are rather frequently used and Instagram bundles popular posts to reach a broader audience. Hence, these data can be considered public if people want them to be public.
The master thesis of Remeeus (2019) provides extra information on the differences between Facebook and Instagram in terms of research and use of data.

3. Data from these social media are collected passively, or without explicit consent.
4. The intended use is to observe persons and enrich the research for one specific client. What is collected precisely should be figured out in a later stage. Data to think about are name, date, description (Marktplaats) or name, relationship status, updates, photos (Facebook) or name, photos, tags, friends (Instagram).
5. The value is the fraud detection, which helps the system of taxes in the Netherlands. Furthermore, the model should be reliable, but also fair. The model should also provide justice to society. An individual whose social media are analysed might feel violated. Or maybe their friend, who are also on some pictures, feel violated since the research was not about them. Hence, this trade-off should be considered.

Luckily, the AP created guidelines to help with concluding this trade-off. The mention that online data gathering can be done on the grounds of covertly perceiving (*nl: Heimelijk waarnemen*)[44]. Hence, this may only be done when there already is suspicion. They specifically mention that it is not allowed to preventively consult these website as it is disproportional to the individuals. This means that it is not allowed to use social media data on a larger group of customers.

So, it seems logical to assume that the data could be enriched by adding certain data to it. Data that could be thought of are social media (including Marktplaats), Kadaster, Belastingdienst, RDW, bank statements, GBA and customer managers/participation process mentor observations. All of these, except for the last two, could be added automatically from a technical perspective. From a legal perspective all of these can only be added if there is reasonable suspicion about someone. If, in the future, this would change, it might be good to start with adding Marktplaats or Instagram data, from an ethical perspective. In this case one might want to start with the Marktplaats data, because of the openness of the data.

---

44 https://autoriteitpersoonsgegevens.nl/sites/default/files/atoms/files/protocol_internetonderzoek_door_gemeenten.pdf

However, Marktplaats can be used by people under a username, while people on Instagram often use their own name. Hence, it might be more feasible to start with Instagram, even though it is more private. From an accuracy perspective there seems to be potential in using additional data, however currently it does not seem like the right place and time for adding these.

7.4 Legitimacy

Legitimacy can be described as legal or lawful. While part of this has already been described in sub question 2. In Appendix II of the SUWI it is described what data municipalities are allowed to store and use about customers. These are the following:

- Data on dates social security assistance
- Data on measure
- Data on housing
- Data on living arrangements
- Data on the standard payment (*nl: normbedrag*)
- Data on termination and claims
- Data on disbursements
- Data on special assistance
- Data on income, classification of the Besluit bijstandsverlening zelfstandigen 2004 (self-employed persons decree) and currency
- Data on social security assistance
- Data on termination and claims
- Municipal data on reintegration
- Data on target group
- Data on participation process (*nl: trajectplan*)
- Data on wage subsidy
- Data on exemption from work obligation
- Data on participation place
- Details of reintegration position
- Data on benefit status

As can be seen, these categories are rather broad; it is not exactly clear what 'data on special assistance', for example, exactly entails. This gives some freedom, but it also means that all data used in the algorithm are indeed legal to store. Since the data may be stored by the municipality itself, they are also allowed to use these data for an algorithm that helps them to more efficiently detect fraudsters. The purpose of storing and using these data must be documented properly such that the data can be used, as was discussed in the purpose limitation part of sub question 2. This may be stated as 'Helping citizens who need social security assistance', for example. If the goal is stated too narrowly, the data may not be used, because of the purpose limitation described in the GDPR.

There are some parts about the data, which should be considered, however. For example, saving a customer's address data is lawfully perfectly fine. This address is needed for sending mail and for house visits, amongst others. However, adding these data to the algorithm might create the possibility of bias based on neighbourhood. The same goes for ethnicity.

This goes towards the line between legality and ethics, because no clear rules are present about this yet. The GDPR states that personal data must be processed fairly, for example, but fair is a subjective term. Hence, these will be discussed under 7.5.

In short, all data used for the algorithm are data that may be stored by the municipality according to the SUWI decree Appendix II[45]. These defined data are rather broad, however. Data stored by the municipality may be used for fraud detection research. Some data, however are questionable to add to an algorithm, because they might introduce a bias. These will be discussed in 7.5, because of their subjective nature.

7.5 Justness

Although the framework has been filled in earlier, in the ethical part of this thesis, this part specifically looks into the ethical side of the attributes used in the algorithm. Normally, when an organisation would be using the framework in Appendix C, this part would serve as a re-evaluation (with more information being present). For now, to answer this question and to keep the method structured, this will be answered here.

When looking at the framework for attribute-related points, the main focus is to prevent bias and discrimination. Furthermore, it was investigated whether the data used were proportional to the goal. This question was deliberately left for this section to be answered. Next to those, whether the data is sufficiently privacy preserving against attackers with background knowledge or when combined with other data sources was something that came up. Last, whether the algorithm makes efficient use of resources is something that is relevant for the attributes of the algorithm.

7.5.1 The algorithm should be free of bias and discrimination

The algorithm should be free of bias and discrimination and part of this is done by training the algorithm in a proper manner. This training has been discussed in the validation of the framework on Gemeente Amersfoort in Appendix E. Another part is done by using attributes that, to some extent, prevent this from happening. One example of an attribute that might be discriminating was given by the validation of the framework on SyRI in Appendix D. This was address information. With this information it might be that some neighbourhoods are detected by the algorithm as troublesome, purely because of earlier bias of the training data. If it is true, it creates a bias; people from this neighbourhood will more often be classified as potential fraudster because they live there and the algorithm learned there were more fraudsters who lived there.

It is not clear-cut which attributes can create such a bias. Often these are the attributes of which society thinks it would be bad to classify someone on. These can be gender or medical conditions, for example. Attributes in the dataset that could lead to such an unwanted bias could be, according to the author, *gender*, certain categories in blocks containing

[45] https://wetten.overheid.nl/BWBR0013267/2021-01-01#BijlageII

information on *hospital stay* and *detention*, categories in activities (*labour desk*), *integration courses*, *special situation* and *steps*.

The details of *special situation* and the information on *hospital stay* were already not taken into account for the algorithm created in this research, because of the (non-)usefulness of the attribute (*special situation*) and ethical concerns (*hospital stay*). The hypothesis was that the accuracy of the algorithm will go down when removing all other mentioned attributes from the dataset. The dataset used here is the dataset with the attribute selection of 7.2 already done. When testing this hypothesis, the algorithm was run three times again, with a dataset which had all the above-mentioned attributes removed. This average accuracy found here was 69.62%. This accuracy was indeed lower than the attribute selected dataset (70.13%), but higher than the full dataset (69.42 Hence, for the 'cost' of about 0.5% accuracy, most, if not all, ethically questionable attributes could be removed. It is up to the organisation, in this case Gemeente Amersfoort, whether they want to make this trade-off in favour of the ethical side or the accuracy side.

It should be noted that because the accuracy does not go down too much, the information could still be (indirectly) present in the dataset. This happens when a non-sensitive attribute highly correlates with a sensitive attribute and can be used by the algorithm instead. An example of this is that someone who goes on *vacation* quite often (or even vaguer, has had *blocks* quite often) might be of non-Dutch origin.

It needs to be stated here that the RF algorithm in this report certainly differs from the algorithm created by Totta Data Lab. Hence, this process of identifying ethically difficult attributes that might create an (unwanted) bias and possibly removing them should be repeated by them. The ethically questionable attributes do not necessarily need to be removed, since some of them might have a very high predictive power which will lower the accuracy significantly. The trade-off should then also be re-evaluated.

Another interesting point is that the law enforcers can see all the data. This might be questionable, for example, concerning whether someone was in jail or not. The principle of jail time is that when one is released from jail after fully serving their sentence, a person gets a second chance from society. And truly getting a second chance, means that the offence and sentence should not hinder them in finding their place in society again. If this data is stored and used by law enforcers to make decisions upon, this might not be the case. The data, however, is needed to see if someone is eligible for social security assistance . If someone is in jail, they will not receive this money. This dilemma drifts too far from the scope of this research, but it shows that more care is taken when it is about algorithms than humans. This is partly sensible, since algorithm would use way more data than humans possibly could use. Whether all these data are needed and/or wanted for the law enforcers is something municipalities in general should question.

7.5.2 The data that is used is proportional to the goal

As mentioned during the validation of the ethical framework in 6.4 the goal is quite large; having only legitimate social security assistance, hence needing to detect fraudsters. As also mentioned, the amount of money that goes with this fraud is also a huge sum. Because the

goal is so big and important, quite some data may be used for it to be proportional. It was already concluded that all data may be stored and used by the municipality for fraud detection. Especially with the ethically questionable data removed, the data that are left seem reasonable for a fraud detection research. Removing data that did not add much also makes that the data that are left are more useful. Even the full, original dataset seems, in terms of data, to be fairly proportional to the goal, mainly because no other data sources are coupled.

7.5.3 The data should also be relatively privacy friendly for people with background knowledge or when combined with other data sources

This question was already answered in the validation of the ethical framework on Gemeente Amersfoort in Appendix E, Non-Maleficence. It was stated that all data are pseudonymized, but are still identifiable with correct background information. This should not be a problem, since the people who have access signed a processing agreement. This conclusion does not change based on this part.

7.5.4 The algorithm should make efficient use of resources

In terms of money, the same goes as was discussed in Appendix E, Justice. Still one fraud case needs to be detected by the algorithm annually for it to break even in terms of cost. In terms of energy, the main gist is still the same, however, with the attribute selection, the runtime of the algorithm was reduced by around 43%. After filtering out the ethically questionable attributes, the runtime was even further reduced to around 18 minutes, or around 50% compared to the original runtime. Hence, it can be useful to look closely at the attributes to see which ones are really needed, to optimize the energy usage of the algorithm. It should be noted here that these runtimes are still very short, perhaps even negligible. This still improves upon humans, who consume quite some energy, likely more than an algorithm would. In terms of manpower, as also discussed in Appendix E, Justice, there is still more manpower needed with the algorithm than without, since people are involved with the development and maintenance of the algorithm. On the other hand, the law enforcers should need less manpower, because they fish in the pond with the suspicious fishes. In terms of data, it is still about equally as efficient. The algorithm uses less data than the law enforcers, because it gets a subset. After filtering a part out because of the attribute selection and because of ethical concerns, this subset became even smaller. It is assumable that Totta Data Lab will do something similar.

7.5.5 Conclusion

Four questions of the framework that were specifically about attributes were answered here. First, it was found that removing all ethically questionable data from the dataset lowered the accuracy by 0.5%. This means that a decision should be made by the organisation on whether this is worth it. This, of course, differs per organisation and should be evaluated on per case. Even the algorithm in this research differs from Totta Data Lab's algorithm, thus should be evaluated as well. However, this research indicates that the possibility is there. Second, it seems that the amount of data used is fairly proportional to the goal. Third, the opinion on whether the data are preserving against people with

background knowledge did not change. Fourth, the algorithm makes reasonably efficient use of resources, especially after attribute selection and when outsourcing it (and deploying it in the cloud).

7.6 Implications for the design of fraud detection algorithms

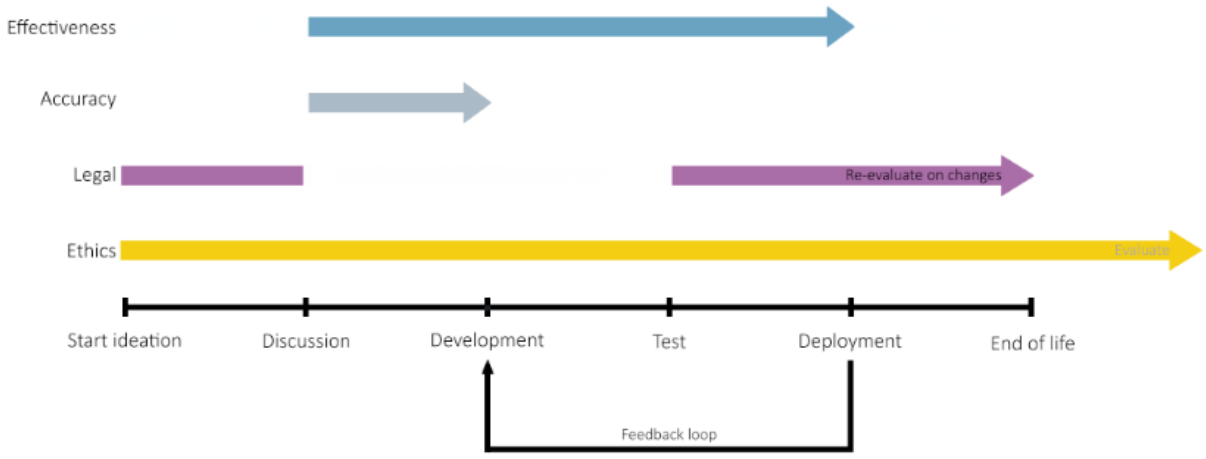


Figure 24 - Implications of the research on effectiveness, accuracy, legitimacy and just on the designing phase of fraud detection algorithms

Much of what was done in this sub question should normally be done in or before the development phase of an algorithm. Some steps being seemingly more logical than others. For example, attribute selection (effectiveness) and potentially adding data sources (accuracy) are often already done. Making sure all data are legal to use and store is something that is done in most cases as well. The ethical part, however, is something that is often overlooked, as also mentioned by a report on algorithms by de Rekenkamer (2021). The main takeaway from this sub question in regards to designing such an algorithm was that the ethical part is intertwined during the entire process, where it should be started the earliest and ended the latest and updated in all phases in between. A summary of this conclusion can be found in Figure 24.

A project starts most often with an idea. Within this ideation phase, it is good to start with the ethics here already. Handles for this are given in Appendix C. Especially the broader question can be answered here already. For example, whether the idea suits the organisation. In this phase too, it is important to consider whether the idea is feasible in terms of legal aspects. A legal framework for this was developed in this research, which can be found in Chapter 4, specifically in Figure 11. It was found in this question that municipalities are allowed to use and store many data. Which data specifically are specified in decree SUWI. There is not necessarily a difference between the permission of having and using certain data with or without an algorithm. The difference lies in that external sources may not be automatically processed, but may only be individually requested when suspicion is present. The algorithm can thus never use these data in an automated way, while humans can in certain cases. This may change in the future, and thus should be re-evaluated.

After this, the phase continues in discussion and further ideation. It is important to detect here which data will be used and why and whether there should be data added or removed. This must be based on the legal part and on the ongoing ethical framework. Like in every phase, the ethical framework should be updated with the newly found information. In this way, the ethical framework will become more and more complete over time, but it also provides certain guidelines during the process.

When the algorithm is really being developed, it becomes interesting to see which attributes might be questionable from an ethical perspective. One way to do this would be to sit together with the team (which may or may not include a data ethicist) and let the employees categorise the different attributes on how privacy-intrusive these are. For the high privacy-intrusive attributes, or attributes the employees have other (ethical) concerns about, it might be good to see how much they actually add to the decision-making process and to maybe leave them out instead. Important is that not too many attributes can be removed, because this might introduce a bias or it might drop the accuracy. On the other hand, is it also important to not invade people's privacy too much by leaving in questionable attributes. Hence, there is a trade-off. For the effectiveness and justified part, it was indicated that there may indeed be attributes that either do not add something or are questionable and removing that does not drop the accuracy that much, but substantially improves the privacy aspect and runtime. Hence in this part, the effectiveness and the ethical side go hand in hand. Furthermore, the ethical framework must be updated with the newly found information. It might feel strange to the reader that accuracy is not incorporated in this phase or the next one. This is because adding data sources can change the algorithm in such a way that it can be considered a new algorithm. Hence, this should restart the entire process. It is thus important to note here that the process can be displayed as a cycle. When big changes are going to be implemented, the process starts over. For simplicity of visualisation, the process is here displayed as linear.

Next, the algorithm might go to the testing phase where it would a great idea to double check on the legal part and the effectiveness part. It might be that during testing new information is gained or that during this phase the public is being informed and raises questions about this. This may lead to reconsiderations on the ethical or effectiveness part. The ethical part is also here ongoing, because of new information that most likely will be gained.

After this, the algorithm may go toward deployment. At this stage, the algorithm should be regularly re-evaluated on the ethical aspect. Again, if drastic changes need to be made, the entire process should start over, since it can be considered a new algorithm (for example adding a new external data source or extreme public outrage about this or a similar algorithm). If laws are altered, the legal part should be re-evaluated as well.

After the End-of-Life phase, it might be good to evaluate back upon with new knowledge. This should be used to learn from as an organisation. An example of this is that lessons should be learned from SyRI and the way it was implemented to prevent the mistakes made in later implementations.

This process is described in Figure 24. For the exact steps within certain phases, the abovementioned text should be consulted. It might be that an organisation wants to continue a phase later on or start a phase earlier or there might be important changes somewhere during the process. The figure serves as an indication. The algorithm may stop being developed during any phase for any reason, for example a lack of funding.

7.7 Conclusion

The goal of this part was to answer the sub question: ' Are the conventional assessment criteria of the algorithm legitimate, justified, effective and accurate? (And what are the implications on the designing process?)'.

For effectiveness, first an attribute selection was performed where attributes that did not add much to the decision-making process were removed or recoded, as well as highly correlated attributes. In total 43 attributes or categories were removed or recoded. The accuracy of the RF algorithm with the full dataset was 69.42%, while the accuracy of the RF algorithm with the altered dataset was 70.13%. The accuracy was thus slightly raised, but rather steady overall. This also reduced the runtime by more than 40%. This gives the indication that extra care should be taken on whether all attributes must be added to the algorithm.

Second, the hypothesis that the accuracy would be raised by adding extra data sources, was tested for the accuracy part. The external sources considered were either social media sites or other instances. Data sources from other sources cannot be coupled because of the GDPR and the current sentiment of society about coupling these data after SyRI. Social media can also not be coupled, because of a document published by the AP. Suppose, in the future it is allowed to use these sources, the author believes it would be good to start testing whether the accuracy would indeed go up with Marktplaats or Instagram, because of their relatively low privacy impact.

Third, for legitimacy, it was found that municipalities are indeed allowed to store and use all data used in the algorithm, as long as the goal is properly defined. Whether all specific attributes may be used cannot be said with full certainty, since the defined categories that may be stored and used by the municipality are rather broad. It was also found that the legitimacy part goes partly hand in hand with the ethical part.

Fourth, the ethical side was looked into. Four remaining questions of the ethical framework, displayed in Appendix E, which were related to the attributes specifically were answered or explored further. This serves as an example of the ethics being present through the entire process. First, all ethically questionable attributes were removed, and the result was that the accuracy went down by about 0.5%. This trade-off between the accuracy and privacy must be considered by the organisation. Furthermore, the ethically questionable attributes were now determined by the author in this case. It would be an idea to have a session with the entire team (including possibly a data ethicist) where the team categorises all attributes according to whether they think they can be used without problems in the algorithm. Second, the data seems proportional to the goal, because the goal is fairly important. Especially after the effectiveness and the removal of ethically questionable attributes. Third,

people with background knowledge still are able to recognise certain cases, but these people are trusted and have signed an agreement. Fourth, the algorithm makes reasonably efficient use of resources.

The implications on the designing process are that it was identified which of the four parts above to focus on during which phase. Furthermore, it was found that ethics play a role during the entire phase, even at the very beginning and after the End-of-Life moment of such an algorithm. A graphical representation of this can be found in Figure 24.

In short, the answer on this sub question is that the assessment criteria can be legitimate, justified, effective and accurate. This cannot be said with full certainty, because much of the legal part is a relatively grey area and because it partly depends on the (ethical) decisions of the organisation. Indications were found of the possibility of removing ethically questionable attributes without losing too much accuracy, which improves the justified part drastically. Detecting these questionable attributes might be something that should be done by the entire team. The main implication on the designing process is that the ethical side should be intertwined in the entire process.

Chapter 8

8. Application to Gemeente Amersfoort

In this chapter, the results of the previous sub questions are applied to Gemeente Amersfoort.

8.1 Method

Most sub questions were focused on formulating a broad answer, such that the result of the sub question could be used for other cases as well. All questions have been answered with the help of information from Gemeente Amersfoort, either by interviews or the dataset they provided for this research. The conclusions of the sub questions, however, were generic. In this question, the conclusion of the previous five sub questions will be tailored to the situation of Gemeente Amersfoort. This sub question is mostly a summary of the previous sub questions. It mentions where the answers can be found in the text above and it applied the legal framework to the case of Gemeente Amersfoort.

8.2 What is the current situation regarding fraud detection at municipalities?

The exact situation for Gemeente Amersfoort specifically can be found in Figure 10. Gemeente Amersfoort differs from other municipalities in that it is strict in the beginning of the process and fairly social during the process. If an algorithm were to be implemented, it can provide for a new incoming signal.

8.3 What criteria of fraud detection algorithms make them lawful to be used?

The criteria that make these algorithms lawful are the same for Gemeente Amersfoort as for other municipalities. Whether they indeed comply with the legal focus points will be looked at here. For this, Figure 11 will be followed from bottom to top. This figure consists of four parts. First of all, the ethical side was already dealt with in this report, as stated below, in the next sub part.

Second is about what needs to be checked when having no algorithm, so in the current situation this should be answered as well. The necessity is there, because social security fraud is a large issue in society. Purpose limitation can be there if the goal is properly defined. The goals should be defined as something like ‘helping all customers in the best way possible’ or ‘only providing rightful social security assistance’. Currently, it is not exactly sure how the goal is formulated. The data minimization part is fulfilled, especially after the attribute selection. It is advisable that Totta Data Lab would also perform an attribute selection. Without the attribute selection, it would probably also be fulfilled, since Gemeente Amersfoort is allowed to store and use all data in the current dataset. However, in the algorithm of this research it was shown that performance could be reached with less data as well. The data are up to date because customers have the duty to inform the municipality on changes. Whether data are secure cannot be said for sure, but it can be assumed it is reasonably secure. This is something that can be found in the DPIA that was conducted. Protocols for safe data handling are in place. Data are deleted after 5 years if no longer needed, this is not easily found on the website, however. It would be good to further specify this, such that people known their data might still be used 5 years later (also to combat fraud and, in this case, train an algorithm to do so). Customers can know which data

are used either by the SUWI decree or by the register of processing operations of Gemeente Amersfoort[46]. Last, as mentioned before, a DPIA had been conducted.

Third, since Gemeente Amersfoort is working towards using an algorithm, the static algorithm part of the legal framework must be looked at as well. Gemeente Amersfoort is working with Totta Data Lab and one of the parts they were working on was on explainable AI. Currently, it would be possible to explain post-hoc why someone was considered a risk case, but it should then be easier to do so. Currently, this research has not been finished yet. Thus, transparency of the algorithm is present to some extent, and will be improved upon to a certain extent. This does not mean that every customer must know exactly what the algorithm checks upon, because this might create circumvention opportunities. The algorithm does decide who may be considered risk cases, but the law enforcers decide whether further research is needed and, in the end, whether someone indeed committed fraud or not. It is thus a uncertain whether this falls under no automatic decisions are made. On one hand, a human is always present in the last steps of the process. On the other hand, customers are still impacted, based on the decisions made by the algorithm. It is also a bit unsure whether, if it falls under automatic decision-making, it is justified as an exception or not. This could fall under national safety or preventing illegal acts, for example. The SyRI case did also not judge on this part. A risk indication by the algorithm does not always lead to intensive research, however. False positives can be filtered out. Gemeente Amersfoort should make a decision on this, since there is no clear statement on this yet. Customers are informed when research on them is being performed, because they are always invited for a conversation. Creating a (completely) fair algorithm is a difficult task, because algorithms are created by humans and, in this case, trained on human data. Humans always (unknowingly) have bias in their decision-making, which means that the algorithm has too, since it is based on data (and knowledge) of humans. This should be trained out over time, by providing the algorithm feedback on its decisions. Furthermore, as described in sub question 5, it should be considered whether all stigmatising attributes are needed in the algorithm. This is also closely related to having a high accuracy. This depends on the trade-off again. Part of the accuracy point is already tackled by letting law enforcers perform a pre-research where they can filter out the false positives. A decision must be taken on the trade-off between using (potentially) stigmatising data and privacy. This is something focus should be put on. The data are being pseudonymised for Totta Data lab, because it must be possible to know whom belongs to which data record. Full anonymous data is not possible in this case.

Fourth, because the algorithm of Totta Data Lab is a learning algorithm, three extra focus points were mentioned in the framework. First, extra focus should be put on transparency. This is, as mentioned, done by Totta Data Lab and their Explainable AI research. This should be somehow monitored by Gemeente Amersfoort to keep up to date with the latest developments. Without this addition, post-hoc explanation is present already, since it can already be explained why someone was considered a fraud risk afterwards. This could arguably be sufficient, because of the complexity of the algorithm(s) used. Extra focus on the

[46] https://www.amersfoort.nl/bericht/register-van-verwerkingen-amersfoort.htm

purpose limitation and the glossary were mentioned by the lawyer of Gemeente Amersfoort. They were working on this, partly by doing a DPIA, partly by having meetings about this. Last, it is important to know and state clearly that having and analysing these data is necessary to combat fraud. This could be done without an algorithm, but the capacity is limited. Hence, with the goal of having a fair system, using algorithms might be beneficial, because there might be social need.

Again, it is good to state here that the legal side it not set in stone, since most of it is a rather grey area and can be interpreted in multiple ways. No related court case has taken place yet, so until then it depends on interpretation. Furthermore, the framework does not contain every part of every law, but highlights the most important parts for this context. The entire GDPR, for example, still applies and should be obeyed.

In general, Gemeente Amersfoort seems to comply to the legal aspects. Six focus points have been identified here. First, it is not completely sure how the goal exactly is defined. This should be defined such that the use of an algorithm is possible and the goals are compatible. Second, it must be made sure that attribute selection is been performed by Totta Data Lab to make sure no unnecessary data are present in the algorithm as well as to possibly remove stigmatising attributes. This also includes making a decision on the dilemma of privacy versus accuracy. Third, Gemeente Amersfoort should take a stance on whether they feel the way of working with the algorithm indeed does not mean automatic decisions are taken or if so, whether this is justified. Fourth, how long data are stored exactly could not easily be found on the website. Extra information must be added when the algorithm will definitely be implemented, including the exact goal. This could be improved. Fifth, the progress with explainable AI of Totta Data Lab should be monitored regularly to see how this is going. Sixth, it must be checked regularly whether the algorithm is relatively free of bias.

8.4 Which ethical considerations should be taken into account when using a fraud detection algorithm?

Like previous sub question, the ethical considerations are the same for Gemeente Amersfoort as they would be for other municipalities. The ethical framework for this can be found in Figure 21 (short version) and in Appendix C (extended version). The extended version has been filled in for Gemeente Amersfoort specifically in Appendix E. The conclusion to this, including focus points for Gemeente Amersfoort, can be found in the validation of the ethical framework under 6.4.3 Amersfoort.

8.5 In what ways is algorithm-based fraud detection different from human expert-based fraud detection?

This sub question was answered based on information of Gemeente Amersfoort. First a dataset provided by Gemeente Amersfoort was used to see how an algorithm works and later interviews were held with two of the law enforcers of Gemeente Amersfoort. Hence, the results of the sub question will probably apply to many municipalities, since municipalities are quite alike. Gemeente Amersfoort might differ in that they are more social, as earlier mentioned. However, it is known for sure that these results apply to

Amersfoort and therefore, all results of these sub questions, as can be found in 5.5 and in Figure 20.

Interesting to briefly mention here is that the accuracy of the algorithm compared to that of the experts was slightly higher (70% vs 65%). This should be further tested, however, since the experts were only asked for their first intuition and only 10 cases were presented to the law enforcers. It does serve as an indication that using an algorithm can be a good addition.

8.6 Are the conventional assessment criteria of the algorithm legitimate, justified, effective and accurate? What are the implications for the design of fraud detection algorithms?

These questions were also answered with the information and dataset of Gemeente Amersfoort. The answers can be found in 7.7. Even though these answers were made as generic as possible, it was still based on the information provided by Gemeente Amersfoort. Removing attributes that do not add much may raise the accuracy. This should, however, be tested since the algorithm used in this research is not the final algorithm used by Totta Data Lab. The same goes for the removal of ethically questionable attributes. Detecting these attributes is something that still can be done by the team at this point in time and it would be advisable to do so. Furthermore, the ethical framework has been filled in for Gemeente Amersfoort and from here focus points were detected. They should re-evaluate this framework periodically and when going to the next phase of the project. Up until now they loosely followed the timeline displayed in Figure 24. It is up to Gemeente Amersfoort to continue this, especially the ethical side.

8.7 Conclusion

As mentioned earlier, this question was mainly a summary of the previous sub questions. It was described where the implications of these sub question for Gemeente Amersfoort could be found in the report. In addition, new information was found for three sub questions. First, in applying sub questions 2, five focus points have been discovered by applying the legal framework to the case of Gemeente Amersfoort. These focus points were purpose limitation, data minimization, (no) automatic decision-making, communication, transparency and fairness of the algorithm. Second, for sub question 4 it was determined that there might indeed be potential in using an algorithm as an addition to the law enforcers, because of the possible gain in accuracy. Third, indications for extra data minimization, also from an ethical perspective, were found. Furthermore, it would be good if Gemeente Amersfoort would continue working on the ethical aspects by involving team members and re-evaluations periodically and per new phase.

Chapter 9

9. Concluding Remarks

This chapter serves to conclude this research. The practical and scientific contributions are discussed here.

9.1 Research questions

This thesis investigated how data-driven fraud risk assessment algorithms can be designed such that they are legally privacy-proof and ethically justified. By answering the sub questions summarized below two frameworks were designed with focus points addressing the ethical and legal side. A timeline was provided as a supplement to the frameworks. Following this timeline at minimum should ensure the right legal and ethical steps were considered throughout all design and implementation phases of deploying fraud detection algorithms.

1. What is the current situation regarding fraud detection at municipalities?

The first research question was answered by performing semi-structured interviews at Gemeente Amersfoort. The fraud detection process without an algorithm was discovered in Chapter 4, specifically in Figure 10. Other municipalities might differ from this figure; however, the big lines will be similar.

2. What criteria of fraud detection algorithms make them lawful to be used?

This question was answered by informing related laws and going through similar cases. Furthermore, an interview was held with the lawyer of Gemeente Amersfoort. From this information, a framework has been created which provides handles for the legal side of creating and implementing these fraud detection algorithms. This conclusion can be found in 4.5, specifically in Figure 11. It was found that the handles differed depending on whether an (learning) algorithm that was used. Furthermore, it was found that the main focus points for a learning algorithm were on transparency, purpose limitation and necessity.

3. In what ways is algorithm-based fraud detection different from human expert-based fraud detection?

Question three was answered by creating a similar algorithm based on public information of Totta Data Lab, supplier and maintainer of Gemeente Amersfoort's and other municipalities' fraud detection algorithms. This was then compared to semi-structured interviews conducted with two of the law enforcers (experts) at Gemeente Amersfoort. Eight differences were found and described in 5.6, specifically in Figure 20. In general, expert-based fraud detection was done by creating personas and looking at the context, whereas the algorithm did not use these more social data.

4. Which ethical considerations should be taken into account when using a fraud detection algorithm?

To answer this question, literature research has been performed in search of ethical frameworks regarding algorithms and a brainstorm has been held. These results were combined into one ethical framework. The framework was divided into five parts: beneficence, non-maleficence, autonomy, justice and explicability. A short, compact version has been made of this framework and an extended version was also created to be used like a checklist. The results provide a solid framework for handles regarding the ethical side of developing and implementing fraud detection algorithms. The framework was validated by applying it to two cases, that of SyRI and Gemeente Amersfoort. The results can be found in

6.5, the (short) framework can be found in Figure 21 and the extended version of the framework can be found in Appendix C.

5. Are the conventional assessment criteria of the algorithm legitimate, justified, effective and accurate?
    a. What are the implications for the design of fraud detection algorithms?

This question was answered by assessing all four criteria step by step. The legal part looked into whether it is legal to store and use the current assessment criteria. The answer to that was that it is legal to store and use the data, depending on the interpretations of the legal part. For the ethical part, the specific questions within the framework on the attributes were answered. Interesting was that many ethically questionable attributes could be removed without substantially dropping the accuracy. For the effectiveness part, attribute selection had been performed in which 20% of the total attributes could be removed or recoded. For the accuracy, it was investigated whether extra data could be added. In this place and time this is not allowed. If it were to be allowed in the future, a start could be made by adding certain social media sites. The general answer to this question was that the conventional assessment criteria can be legitimate, justified, effective and accurate, depending on some legal interpretations and handling of the ethical side. The full conclusion of this can be found in 7.7.

A timeline has been created to answer the second part of this sub question. The main takeaway here was that the ethical part should be intertwined in the entire process of the algorithm: from idea to after end-of-life. This timeline can be found in Figure 24.

6. What do the results of question 1-5 mean for Gemeente Amersfoort specifically?

The implications of the sub questions for Gemeente Amersfoort were mainly that they are given clear focus points to work on both ethical and legal aspect (sub question 2, 4 and 6). Furthermore, their fraud detection process without an algorithm was mapped in sub question 1. Differences between the algorithm and law enforcers were also tailored to Gemeente Amersfoort (sub question 3). Last, work on the ethical and legal side was performed partly for them in this research and can be further used for information provision and ensuring a complete process.

9.2 Discussion and future research directions

This part discusses the method and outcomes of this research. Eight discussion points are mentioned in this chapter, of which 7 can lead to further research.

First of all, some parts of this research did not go as planned because of Corona and the limitations it brought. For example, with sub-question 3. The initial idea was to interview the law enforcers in person, while this was eventually done via an online meeting. This brought complications with it, namely that the law enforcers were not always used to working online, and helping them via an online meeting also proved rather difficult. Other examples are some of the other interviews. When conducting this in person, also non-verbal communication can be picked up by the interviewer. The interviewer could follow up on this and collect more information in doing so. Interviews that were limited in this way were with

the alderman, lawyer, data scientist, and information provision advisor. Therefore, it would be a good idea to do these again, in person, to validate the results of the corresponding sub-question further.

Second, especially the legal framework has been made rather compact. Meaning that there are more lawful parts to comply with than only those described in the framework. However, it is not possible to put all the laws and regulations into one framework, therefore it was chosen to only highlight the most important ones. They should be further tested in practice to further validate this framework and to further test its completeness.

Third, as related to the previous point, the ethical framework has been validated using two case studies. Even though it was deemed a robust tool for detecting ethical focus points in these two cases, the framework should be further tested to further validate the results. Next to that, it would also be interesting to extend the framework on the autonomy part. Currently, the framework assumes subjects do not have a say in whether their data are being used or not, since the data are used for criminal investigation. Interesting would be to extend the framework such that it could also be used for algorithms where subjects can opt-out. This would require an addition to the autonomy part, as was briefly mentioned in the accompanying text. This altered framework should be tested and validated again. Furthermore, it would be interesting to research the possibilities of using the ethical framework in different contexts. The framework will probably need to be altered to fit in that case. However, it seems reasonable to assume that a solid basis is already present in the current framework. Such a different context could for example be in healthcare, where increasingly more algorithms are used, or in e-commerce. In both these cases, ethics seem to become gradually more important. Further validation and altering of the ethical framework to fit different contexts is something for future research. Next to these, it was found at the very end of this thesis that the ethical framework might benefit from adding some extra, specific, security-related questions under non-maleficence, for example about logging and access rights. These steps were implicated in the accompanying test, but it might be good to add these to the framework explicitly as well.

Fourth, the Algemene Rekenkamer (2021) published a report on ethics and algorithms at the very end of this research. It would be very interesting to combine their findings with the ethical framework developed in this research. It would probably create an even more complete picture. The Algemene Rekenkamer did not apply their framework to case studies, however, they did analyse case studies. It would be good to apply the framework of the Algemene Rekenkamer to the same studies as done in this research to highlight the differences. Comparing and combining is something that should be researched in future work.

Fifth, it was found in Chapter 5 that the law enforcers focused on different information depending on the type of fraud they suspected someone to possibly have committed. Based on that, it might be interesting to create multiple algorithms for different types of fraud. It would be imaginable that an algorithm for illegal property possession would focus more on vacations, for example. Doing so might increase the accuracy or lead to needing less data (data minimisation). Hence, this would be a very interesting topic for future research.

Sixth, the question arose whether (small) municipalities are recommended to use an algorithm like the one of Totta Data Lab based on this research. That is difficult to say based because no long-term tests have been conducted yet. This research looked at the ethical and legal side of developing and implementing such an algorithm. With the right interpretations, preparations, and thoroughness of the organisation, this can be arranged. From the first short term results of Totta Data Lab (as mentioned during meetings) it was cautiously noted that the algorithm detected more fraudsters at the start of the implementation and that this decreased over time. It is too soon to state this with certainty. The positive side is that the algorithm needs to detect one fraud case per year for it to break-even cost-wise. Hence, short-term it seems very valuable for (small) municipalities to make use of a fraud-detection algorithm. Long-term success is unsure as of now and is something that should be researched, just as the questions whether there are differences between using a fraud-detection algorithm for large municipalities and small municipalities.

Seventh, law enforcers can currently see and use many data, way more than the algorithm can. To some extent, it is strange that there are such strict rules for algorithms, while these are not present for humans. There is, of course, a difference in scale, and therefore in potential privacy infringement, present. It might be interesting to take a good ethical look into whether this is desirable and how to communicate this. This was beyond the scope of this project; however, it would be an interesting future research topic.

Eighth, the reader might question paying for a solution like the algorithm of Totta Data Lab, while an algorithm was created in this thesis. This indicates the possibility for an organisation to create such an algorithm themselves. It is indeed doable to not outsource this. However, there are good reasons to outsource this as well. First of all, the price per year is equal to stopping one unfairly received social security assistance. So as long as the algorithm finds one extra fraudster per year, cost-wise it is feasible. Second, it is good to not have to spend the, often limited, capacity of municipalities on creating, using, updating, and documenting this algorithm. Third, it is now very clear that Totta Data Lab is responsible for when something about the algorithm does not work. With the contract, the responsibility of good results lies in their hands. Fourth, they have more time to tweak and optimize the algorithm in all areas, especially when multiple municipalities purchase the same service. Whether an in-house or an outsourced solution is preferred is something that should be researched per case.

9.3 Scientific contribution

Many different ethical frameworks were already present, with their focus on specific aspects of the process (data usage or communication, for example). These frameworks have been combined into one solid ethical framework tailored to the context of fraud detection algorithms at municipalities. This framework is not created for a Subpart of the process of implementing an algorithm, but for the entire process; from data to protocols. It aims to close the gap between IT and business on ethical aspects for fraud detection algorithms in the context of municipalities. With some adjustments, the framework could be tailored to other contexts, thereby serving as a solid basis. Furthermore, insights were found on the relatively new process of fraud detection with the use of algorithms. These insights are not

only on the legal and ethical parts, but also the general process. Municipalities can use these insights to support and strengthen their design and implementation phase. Last, the utter importance of thoroughly considering the ethical aspects throughout the entire process was something discovered and emphasized during this research.

9.4 Practical contribution

On the practical side, Gemeente Amersfoort has a thorough analysis of the ethical and legal aspects of their fraud-detection algorithm and the surrounding process as conducted in this research. Furthermore, the ethical and legal frameworks can be used in practice by other municipalities. With some adjustments, the context could be extended. Both will lead to a better and more complete understanding of the topic of fraud detection with the use of an algorithm, for Gemeente Amersfoort specifically and other municipalities if they use the frameworks.

References

Algemene Rekenkamer. 2021. ‘Aandacht voor algoritmes’. https://www.rekenkamer.nl/publicaties/rapporten/2021/01/26/aandacht-voor-algoritmes (January 27, 2021).

Baxter, Gordon, and Ian Sommerville. 2011. ‘Socio-Technical Systems: From Design Methods to Systems Engineering’. *Interacting with Computers* 23(1): 4–17.

de Raad voor de Volksgezondheid en Zorg. 1999. ‘Ethiek met beleid’. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjsnpmJuOjrAhXsIMUKHeIlDWEQFjAIegQICRAB&url=https%3A%2F%2Fwww.raadrvs.nl%2Fbinaries%2Fraadrvs%2Fdocumenten%2Fpublicaties%2F1999%2F12%2F06%2Fethiek-met-beleid%2FEthiek_met_beleid.pdf&usg=AOvVaw33o14BY5oRFrknPsuLVfwn (September 14, 2020).

Doove and Otten. 2018. ‘Verkennend onderzoek naar het gebruik van algoritmen binnen overheidsorganisaties’. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwidmLCBz7fnAhXHIVAKHaBLB8UQFjAAegQIARAB&url=https%3A%2F%2Fwww.cbs.nl%2F-%2Fmedia%2F_pdf%2F2018%2F48%2Fverkennend%2520onderzoek%2520naar%2520het%2520gebruik%2520van%2520algoritmen%2520binnen%2520overheidsorganisaties.pdf&usg=AOvVaw0ckz5yVYArfCziGKIJLcDh (February 4, 2020).

ECP. 2018. ‘Artificial Intelligence Impact Assessment’. https://ecp.nl/wp-content/uploads/2018/11/Artificial-Intelligence-Impact-Assesment.pdf (February 4, 2020).

European Group on Ethics in Science and New Technologies. 2018. ‘Statement on Artificial Intelligence, Robotics and “autonomous” Systems’. https://ec.europa.eu/info/sites/info/files/european_group_on_ethics_ege/ege_ai_statement_2018.pdf (February 7, 2020).

Floridi, Luciano et al. 2018. ‘AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations’. *Minds and Machines* 28(4): 689–707.

Geldrop, Alex van, and Theo de Vries. 2015. *Fraude loont ... nog steeds*. SST. https://research.utwente.nl/en/publications/fraude-loont-nog-steeds (February 12, 2021).

Goebel, Randy et al. 2018. ‘Explainable AI: The New 42?’ In *Machine Learning and Knowledge Extraction*, Lecture Notes in Computer Science, eds. Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl. Cham: Springer International Publishing, 295–303.

Hill, Clara E. et al. 2005. ‘Consensual Qualitative Research: An Update’. *Journal of Counseling Psychology* 52(2): 196–205.

Holzinger, Andreas, Peter Kieseberg, Edgar Weippl, and A. Min Tjoa. 2018. 'Current Advances, Trends and Challenges of Machine Learning and Knowledge Extraction: From Machine Learning to Explainable AI'. In *Machine Learning and Knowledge Extraction*, Lecture Notes in Computer Science, eds. Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl. Cham: Springer International Publishing, 1–8.

Ipsos. 2018. 'Kennis Verplichtingen en Detectiekans 2017'. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=2ahUKEwiynebd5rfnAhVFIlAKHdzcB8UQFjABegQIARAB&url=https%3A%2F%2Fwww.rijksoverheid.nl%2Fbinaries%2Frijksoverheid%2Fdocumenten%2Frapporten%2F2018%2F05%2F30%2Fkennis-verplichtingen-en-detectiekans-2017%2F17039811_Ipsos_rapport_Kennis%2BVerplichtingen%2Ben%2BDetectiekans%2B2017.pdf&usg=AOvVaw0xB1g70Y4uz39aP_qAMFi6 (February 4, 2020).

Kimbrough, Steven O., D. J. Wu, and Fang Zhong. 2002. 'Computers Play the Beer Game: Can Artificial Agents Manage Supply Chains?' *Decision Support Systems* 33(3): 323–33.

Knox, Sarah, and Alan W. Burkard. 2009. 'Qualitative Research Interviews'. *Psychotherapy Research* 19(4–5): 566–75.

Leenes, Ronald. 2016. 'De voorspellende overheid: Transparantie is noodzakelijk, maar hoe?' *Bestuurskunde* 25(1). http://www.boomuitgeversdenhaag.nl/home (February 11, 2020).

Lipton, Zachary C. 2018. 'The Mythos of Model Interpretability'. *Communications of the ACM* 61(10): 36–43.

Ministerie van Justitie en Veiligheid. 2019. 'Kamerbrief over waarborgen tegen risico's van data-analyses door de overheid - Kamerstuk - Rijksoverheid.nl'. https://www.rijksoverheid.nl/documenten/kamerstukken/2019/10/08/tk-waarborgen-tegen-risico-s-van-data-analyses-door-de-overheid (May 1, 2020).

Olsthoorn, Peter. 2016. *Big Data voor Fraudebestrijding*. Den Haag: Wetenschappelijke Raad voor het Regeringsbeleid. www.wrr.nl (January 18, 2020).

Peffers, Ken, Tuure Tuunanen, Marcus A. Rothenberger, and Samir Chatterjee. 2007. 'A Design Science Research Methodology for Information Systems Research'. *Journal of Management Information Systems* 24(3): 45–77.

Raat, Caroline. 2020. 'WET-SYRI IS ONRECHTMATIGE OVERHEIDSDAAD'.

Rai, Arun. 2020. 'Explainable AI: From Black Box to Glass Box'. *Journal of the Academy of Marketing Science* 48(1): 137–41.

Reddix-Smalls, Brenda. 2011. 'Credit Scoring and Trade Secrecy: An Algorithmic Quagmire or How the Lack of Transparency in Complex Financial Models Scuttled the Finance Market'. *UC Davis Business Law Journal* 12: 87.

Reinsel, David, John Gantz, and John Rydning. 2018. 'The Digitization of the World from Edge to Core'. : 28.

Remeeus, Lindy. 2019. 'Participant Perceptions of Facebook and Instagram Research Ethics: A Replication Study.' Tilburg University. http://arno.uvt.nl/show.cgi?fid=150110 (January 18, 2020).

Rowley, Jennifer. 2012. 'Conducting Research Interviews'. *Management Research Review* 35(3/4): 260–71.

Shubhendu, Shukla S. and Vijay, Jaiswal. 2013. 'Applicability of Artificial Intelligence in Different Fields of Life'. *International Journal of Scientific Engineering and Research (IJSER)* 1(1): 8.

The IEEE Global Initiative on Ethics of Autonomous and Intelligent System. 2019. 'Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems, First Edition'. *IEEE*. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.htm.

Totta Data Lab. 2020. 'Technische documentatie voorspelmodel bijstandsfraude gemeente Nissewaard'. https://nissewaard.notubiz.nl/document/8567516/1/Beantwoording_LOB%2C_SyRi-uitspraak_voor_algoritme_Totta_Data_Lab%2C_bijlage (October 2, 2020).

van der Weerdt and de Vries. 2014. 'Dienstverlening Verbeteren met Big Data'. https://kennisopenbaarbestuur.nl/media/62989/dienstverlening-verbeteren-met-big-data.pdf (February 4, 2020).

van Veenstra et al. 2019. 'Quick Scan AI in de Publieke Dienstverlening'. https://www.rijksoverheid.nl/documenten/rapporten/2019/04/08/quick-scan-in-de-publieke-dienstverlening.

Wieringa, Roel J. 2014. *Design Science Methodology for Information Systems and Software Engineering*. Berlin Heidelberg: Springer-Verlag. https://www.springer.com/gp/book/9783662438381 (September 18, 2020).

van Wynsberghe, Aimee, Henry Been, and Maurice van Keulen. 2013. 'To Use or Not to Use: Guidelines for Researchers Using Data from Online Social Networking Sites'. *Web publication/site, UK: RRI, Rict Responsible Innovation*. https://research.utwente.nl/en/publications/to-use-or-not-to-use-guidelines-for-researchers-using-data-from-o (February 5, 2020).

Xu, Feiyu et al. 2019. 'Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges'. In *Tang J., Kan MY., Zhao D., Li S., Zan H. (Eds) Natural Language Processing and Chinese Computing*, Lecture Notes in Computer Science, Springer, Cham. https://link-springer-com.ezproxy2.utwente.nl/chapter/10.1007/978-3-030-32236-6_51.

Zhang, Dongpo. 2018. 'Big Data Security and Privacy Protection'. In Atlantis Press. https://www.atlantis-press.com/proceedings/icmcs-18/25904185 (October 15, 2019).

Appendix

Appendix A

Below are the questions asked to the three stakeholders interviewed in Chapter 3.

1. Introductie
2. Kun je in het kort omschrijven wat je baan precies inhoudt?
3. Hoe gaat het proces van fraudeopsporing in zijn werk?
4. Jullie krijgen wel eens tips binnen van mensen over fraudeurs, om wat voor tips gaat het dan?
    a. Hoe ziet zo'n tip eruit?
    b. Zie je meer tips in een bepaalde categorie (zwartwerken, samenwonen e.d.)?
5. Stel jullie krijgen zo'n tip over een potentieel fraudegeval, wat gebeurt er dan?
    a. Hoeveel procent van deze tips is legitiem?
6. Wat gebeurt er met de mensen waarover jullie geen tip krijgen?
    a. Worden die alsnog doorgelicht?
    b. Hoe gebeurt dit?
    c. Op basis van welke criteria kijken jullie of dit inderdaad een fraudeur is?
    d. Is er een 'prototype' fraudeur? Zo ja, hoe ziet deze eruit?
7. Wordt er actief gewerkt aan het tegengaan van discriminatie?
8. Hoeveel procent van de fraudeurs schat je dat ontdekt wordt?
    a. Hoeveel mensen worden er doorgelicht zonder dat ze fraude plegen?
    b. Hoe denk je dat dit percentage omhoog kan?
9. Weet je of dit ook bij andere gemeenten hetzelfde werkt?
10. Wat denk je zelf van het gebruik van een algoritme om verdachten aan te stippen?
    a. Waar zie je voordelen?
    b. Waar zie je nadelen?
11. Wil je nog iets toevoegen?
12. Afsluiting

Appendix B

The full list of data that can be processed by SyRI, according to Besluit SUWI, article 5a, section 3 is displayed in this appendix.

- work data, being data with which the work performed by a person can be established;

- data on administrative measures and sanctions, being data proving that an administrative fine has been imposed on a natural or legal person, or that another administrative measure has been taken;

- tax data, being data with which the tax obligations of a natural or legal person can be established;

- data on movable and immovable property, being data with which the possession and use of certain property by a natural or legal person can be established;

- data on grounds for exclusion from social assistance benefit or other benefits, being data proving that a person is not eligible for a benefit;

- trade data, being data with which the nature and activities of a legal person can be established;

- housing data, being data with which the actual or other place of residence or place of business of a natural or legal person can be established;

- identifying data, being for a natural person: name, address, city, postal address, date of birth, gender and administrative characteristics, and for a legal person: name, address, postal address, legal form, place of business and administrative characteristics;

- civic integration data, being data with which it can be established if an obligation to participate in a civic integration programme has been imposed on a person;

- compliance data, being data with which the compliance history with legislation and regulations of a natural and legal person can be established;

- education data, being data with which the financial support for the funding of education can be established;

- pension data, being data with which pension entitlements can be established;

- reintegration data, being exclusively the data with which it can be established if reintegration obligations have been imposed on a person and whether or not they are or are being met;

- debt burden data, being data with which any debts of a natural or legal person can be established;

- social benefit, allowances and subsidy data, being data with which the financial support of a natural or legal person can be established;

- permits and exemptions, being data with which it may be established for which activities a natural or legal person has requested or has obtained permission;

- health care insurance data, being exclusively the data with which it can be established if a person is insured for the Healthcare Insurance Act

Appendix C

The full extended ethical framework for the implementation of (fraud) algorithms within public organisations can be found in this appendix.

Benificence

- ☐ Which ethical theory fits your organisation and will be followed throughout this framework?
- ☐ Is the algorithm (generally speaking) beneficial to humanity?
  - ☐ Have you (re-)evaluated laws, human rights and other conventions for this?
- ☐ What are economic, societal and environmental benefits of the algorithm?
  - ☐ Do these benefits fit your organisation and its goals?

Non-Malificence

Privacy

- ☐ Which information (data) are used for the algorithm and why?
  - ☐ Is this information legal to use?
  - ☐ Is this information ethical to use?
- ☐ Is the amount of data used proportional to the goal?
  - ☐ Do all these used criteria add something to the decision-making process?
  - ☐ Is the goal unreachable with less data?
- ☐ Does the algorithm need to be trained?
  - ☐ Can you use a fictional dataset or an anonymous dataset?
  - ☐ Is this dataset representative of the real world?
- ☐ Are data subjects protected from being a false positive (too often)?
  - ☐ Is the impact on someone when they are a false positive acceptable?
- ☐ Is the algorithm also reasonably privacy friendly for people with background knowledge or when combined with other data sources?
- ☐ Will the algorithm still be privacy friendly in 20 years?
  - ☐ Are there people in the team who can imagine a future scenario in which the results of your project can be misused?
- ☐ Have you conducted a DPIA?
- ☐ Have you discussed your ethical ideas and concerns with a privacy officer or an independent organisation?

Misuse

- ☐ Do you have clear protocols regarding what the algorithm may and may not do and how humans should interact with this?

Future

- ☐ Do you have clear time paths for re-evaluating this framework?

Security

- ☐ Is the data secure for the outside and for the inside?

Up to date

- ☐ Will the algorithm be updated regularly?
- ☐ Do you know what 'good behaviour' entails for your algorithm?
  - ☐ How can you check this?
- ☐ Are the data (actively) being kept up to date?
- ☐ Are the data of good quality and well organised?
- ☐ Are all protocols up to date?

Developing

- ☐ Is the algorithm developed by skilful, rational and responsible people?
- ☐ Do you also know the weaknesses of your team?
- ☐ Is the algorithm tested extensively?
- ☐ Is the development documented properly?

Access

- ☐ Is it clear who can access the data and when?
  - ☐ Is this clearly written down, defined and documented?
- ☐ Is there a regulatory body overseeing your algorithm?

Autonomy

- ☐ Is the algorithm unable to autonomously hurt someone?
- ☐ Are there protocols for when humans should interact with the algorithm?
- ☐ Do citizens know how and when they can challenge a decision made by the algorithm?

Justice

Rights

- ☐ Are the rights and interests of all parties analyzed and reflected in the algorithm?

Fair

- ☐ Is the algorithm free of bias and discrimination?
- ☐ Is the algorithm free of any potential discriminating factors?
  - ☐ Do team members believe the algorithm is free of any potential discriminating factors?
- ☐ If there is any unfairness in the algorithm, is this justifiable?

Stakeholders

- ☐ Are stakeholders included in the development process?

Accuracy

- ☐ How accurate is the algorithm?
  - ☐ What is the percentage of false positives marked by the algorithm?
  - ☐ Is there no serious harm of the (in-)accuracy?
- ☐ Is the algorithm adapted to handle new scenarios?

Effectiveness & Efficiency

- ☐ Can your goal be achieved by using an algorithm?
  - ☐ Is this the best way to do so?
  - ☐ How do similar organisations (with similar solutions) perform?
- ☐ Are the results of the algorithm effective?
- ☐ Does the algorithm make efficient use of resources, in terms of money, energy, manpower and data?

Explicability

Stakeholders

- ☐ Have you identified the stakeholders of the algorithm?
  - ☐ Have you also identified indirect stakeholders?
  - ☐ Have you identified their position?
  - ☐ Do you know what their expectations and wishes are?
  - ☐ Is it clear how they are affected?
  - ☐ When will you re-evaluate this?

Context

- ☐ Have you identified the context of the algorithm?
- ☐ Is this unable to change?
  - ☐ When will you re-evaluate this?

Transparency

- ☐ Is it discoverable for the users and data subjects how and why the algorithm made a certain decision?
- ☐ Does your algorithm have full transparency (transparency design) or partially (post-hoc)?
  - ☐ Is it desirable to have full transparency in your algorithm in the context?
- ☐ Can you test the algorithm?
- ☐ Are there protocols for handling data and the algorithm?
  - ☐ Is it transparent which data are used and why, and who can access them at which times?
  - ☐ If the algorithm is malfunctioning or producing unexpected outcomes, is there an exit strategy or change protocol for adjusting the algorithm?
  - ☐ Are there protocols in case of (extreme) public outrage and who is responsible for this?
  - ☐ Are there no other parts in or surrounding the algorithm that need protocols?

Users

- Can the users of the algorithm explain the algorithm?
- Are the users of the algorithm educated on how dependent and reliant they can be on the algorithm?

Education

- Is the public sufficiently educated about the algorithm and do they have a say in it?
  - Do they know the goal of the algorithm, why you chose a certain type of algorithm, which data are used and how often the algorithm will use these data?
  - Do they know what the consequences may be, who is responsible for the analysis and which quality checks are present?
- Is it clear who is responsible for communication with the public?
- Is the communication to the public done in a concise, understandable and easily accessible manner?
- Does your website include the fact that you perform data analysis, why you do this, what the consequences may be for citizens, whether or not you use machine learning and an explanation of this, what the legal basis is, which data sources are used, who is responsible for the analysis, what the role of third parties are in this process, which quality checks are performed, if there is human intervention in the process and which assessment frameworks are present and how they are used?
- Are your employees well-informed?
  - Are their concerns addressed properly?

Citizens

- Are citizens informed when they are checked more extensively by the algorithm or when their data are at risk?
- Do they have the option to raise objections to the decisions made by the algorithm?
- Do you remind citizens of the usage of the algorithm once in a while or at least when something changes?

Accountability

- Is it clear who created the algorithm?
- Is it clear who is responsible for the way the algorithm works, the data that are chosen, what happens when something goes wrong, the usage of the algorithm and for creating protocols?
- Is it clear who is ultimately responsible for the project?
- Are these documented?

Figure 25 - Ethical framework for the implementation of fraud algorithms within public organisations.

Appendix D

This appendix applied the project framework to the case study of SyRI.

Beneficence

✅ Which ethical theory fits your organisation and will be followed throughout this framework?
For SyRI, this is not entirely sure. The government has three tasks according to Ethiek met Beleid (de Raad voor de Volksgezondheid en Zorg 1999). These are ordering, protecting and promoting. This means that the government is a rather social organisation and new algorithms should not harm people. Part of this is utilitarian (most happiness for most people), but also Kantian (people have duties, like working) and a rather new ethical theory: ethics of care (taking care of each other and the world).

✅ Is the algorithm (generally speaking) beneficial to humanity?
Yes, the idea behind SyRI is. This is namely to prevent people from receiving unjust social security assistance and thus making sure the tax money of citizens will be used correctly and creating or containing the trust of the citizens in the government.

☐ Have you (re-)evaluated laws, human rights and other conventions for this?
This was clearly done, since a new law was introduced. However, the law that was introduced was also judged to be unethical. So, the laws are evaluated, but not on ethical grounds.

✅ What are economic, societal and environmental benefits of the algorithm?
Economic benefits are clear, a lot of money is lost due to fraud and SyRI should be able to prevent this. Whether this holds up against the costs for creating the system is unclear. However, this is not the most important goal of SyRI. Societal benefits are that the money of the tax payer goes to the right place. This helps with citizen's trust in the government as well. Environmental benefits are unclear.

✅ Do these benefits fit your organisation and its goals?
Yes, for sure. Government's goals are mentioned above and they relate to the benefits discovered.

Non-Maleficence

Privacy

☒ Which information (data) are used for the algorithm and why?
These can be found in Appendix B of this report. As mentioned by the court in the SyRI case, there are not many personal data to be thought of that cannot be placed under one of these points. The 'why' is of course to catch as many fraudsters, but the clear why per data field lacks.

☑ Is this information legal to use?
They are legal, because the law was created around the algorithm itself.

☒ Is this information ethical to use?
Not really, the data fields are not clearly defined or limited to what is really needed. People cannot know what data about them are really used, which makes it not ethical. There are also some questionable fields included, like education data and gender.

☑ Is the amount of data used proportional to the goal?
It is unsure how much one data field contributes to the percentage of fraudster caught. All data do contribute to the goal; however, it does seem like a whole lot of data and it is unclear why some fields are added. The court judged that the data minimization principle was taken into account, mainly because of the severity of the goal (preventing fraud).

☐ Do all these used criteria add something to the decision-making process?
That is unsure, since not all parts about SyRI are released. However, there are so many data fields and it is not clearly defined which exactly are used. Probably not all the data fields that may be used will add something to the decision-making process.

☐ Is the goal unreachable with less data?

This depends on what the goal of SyRI is. If the goal is to catch all fraudsters, you would probably want as much data as possible. Question is whether this would be a realistic goal. Whether the percentages of people being caught by the algorithm changes drastically by adding more data is unclear. Possibly, the goal can still be achieved with less data.

☑ Does the algorithm work without training
In the SyRI case[28] the state mentioned that SyRI is not a learning algorithm.

☐ Can you use a fictional dataset or an anonymous dataset?
For the algorithm itself, it is not possible to use an anonymous dataset, because the persons must be retraceable. Hence, the maximum amount of anonymity would be with pseudonymisation. This is done for SyRI. The algorithm does not need training. It is unclear with what data SyRI has been developed.

☐ Is this dataset representative of the real world?
It is unclear with what data SyRI has been developed.

☐ Are data subjects protected from being a false positive (too often)?
According to the SyRI case, there is a human action required before creating the final risks selection, where false positives and false negatives should be eliminated from the dataset. It is unclear how this exactly is done.

☒ Is the impact on someone when they are a false positive acceptable?

Court judged that the impact on human life is huge. So huge that it falls within article 8 of the ECHR. This is clear when we look at it, since people who are indeed assigned as a risk are

completely checked; all useful, available data are analysed by the algorithm, without them knowing about this.

☐ Is the algorithm also privacy friendly for people with background knowledge or when combined with other data sources?

Personal names and company names, citizen service numbers and addresses are among the data that are replaced with a code (a pseudonym). The IB (Stichting Inlichtingenbureau) is the organisation who creates the keys for the data sets. The ministry SZW (Social Affairs and Employment) will analyse the decrypted data. So, the IB, SZW, the municipalities and other involved instances all can see (parts of) the decrypted data. SZW gets a limited data set, with only risk cases. Other people should not be able to see the data. Since so many data fields may be used, these pseudonymized data will probably still be recognizable for people with background knowledge or when combined with other sources. In principle, the data set should not be shared, so only trusted people from these instances will be able to see it.

N.B. please note that there was found that this question should be altered. With this alteration, this question would be checked/green. See 6.4.3 Amersfoort for more information.

☑ Will the algorithm still be privacy friendly in 20 years?
Since the datasets are not public, yes it should be the case.

☐ Is there checked whether there are people in the team who can imagine a future scenario in which the results of your project can be misused?

This cannot be discussed in this report since the author does not have a team surrounding SyRI.

☑ Have you conducted a DPIA?

According to the court case, this has been performed.

☑ Have you discussed your ethical ideas and concerns with a privacy officer or an independent organisation?

This is unclear. The LSI (Landelijk Stuurgroep Interventie) was involved, but they have interest in preventing fraud and are therefore not independent. In the Netherlands, several parties are involved when creating or altering a law. These are the Raad van State, who checks the feasibility of the law, the Tweede kamer, who voted on it, and the Eerste kamer, who voted on it again. It is arguable whether these parties have an interest in preventing fraud, since they would also have interest in the privacy of their citizens. Therefore, the author would say that it is discussed with an independent party.

Misuse

☐ Do you have clear protocols regarding what the algorithm may and may not do and how humans should interact with this?

It is rather clear what the algorithm may an may not do. However, the protocols are not public. It is thus unclear whether the protocols are present, let alone complete.

Future

☐ Do you have clear time paths for re-evaluating this framework?

This question cannot be answered, since the people behind SyRi did not fill in this framework at all.

Security

☐ Is the data secure for the outside and for the inside?

For the inside, this cannot be said for sure, since the team involved is unknown to the author. According to the court case, it is secure for the outside.

Up to date

☐ Will the algorithm be updated regularly?

This did not become clear from the court case.

☑ Do you know what 'good behaviour' entails for your algorithm?

It is clear what good behaviour entails.

☐ How can you check this?

It is unclear how this will be checked, since much about SyRI is kept a secret. Court judged SyRI was not sufficiently checkable, but they did not specifically mention that for good behaviour. It is also unclear how verifiable SyRI is for the people working with it, since it was mentioned that a risk case is not reproducible.

☐ Are the data (actively) being kept up to date?

Not much about this is mentioned in the court case. What is mentioned is that citizens are not able to check their own data to see whether they are up to date. The court did mention the importance of up-to-date data, but no follow-up on this is given. It is important, however, to know that SyRI does not create the data, municipalities and other instances do via the citizens and it depends on the organisation how they deal with this point.

☐ Are the data of good quality and well organised?

It is unsure what data are actually used, let alone whether the quality of these data is good.

☐ Are all protocols up to date?

This is unclear.

Developing

☑ Is the algorithm developed by skilful, rational and responsible people?

The system SyRI was created by the government, but it is not clear by whom exactly and whether they are responsible humans. There is a rather extensive procedure before someone is being able to work with the government, however. So, most people should be. Often also a VOG (Verklaring Omtrent Gedrag) is requested, in which is specified whether someone might be a threat to society before getting to work somewhere.

☐ Do you also know the weaknesses of your team?
This cannot be said for sure.

☐ Is the algorithm tested extensively?
Not much information is present about the development phase. Information that is present is about the real-world application of the algorithm. It was only applied to neighbourhoods that are known as problem areas. This might imply that it is not tested extensively, since this does not give a good overview of the real-world situation; not all neighbourhoods are problem areas. It is of course, not sure what was tested behind the scenes.

☐ Is the development documented properly?
This is unknown.

Access
☑ Is it clear who can access the data and when?
As discussed under 'Is the algorithm also privacy friendly for people with background knowledge or when combined with other data sources?', the IB, SZW, the municipalities and other involved instances are able to see the decrypted data. When they are able to do this, is quite clearly described in the court case: 'Data processing takes place in two phases: processing (phase 1) and analysis (phase 2). In the first phase the IB collates the records and pseudonymises them. Personal names and company names, citizen service numbers and addresses are among the data that are replaced with a code (a pseudonym). The processor then applies the first step in the risk selection to the encrypted data: the source file is checked against the risk model with all indicators in an automated manner. This generates potential hits. A potential hit is a hit that indicates an increased risk of fraud. The IB also creates a key file specifying which personal name or company name, citizen service number or address belongs to which pseudonym. When based on the risk model certain natural persons, legal persons or addresses are flagged as an increased risk, they are decrypted with the key file. All data related to these increased risks, except for the key file, are then forwarded to the Minister for the second phase of risk analysis by the analysis unit of the Social Affairs and Employment Inspectorate. The IB destroys any SyRI project files still in its possession within four weeks from forwarding the data to the Minister. The destruction is laid down in an official report.

In the second phase the decrypted data are analysed more closely by the analysis unit of the Social Affairs and Employment Inspectorate. The data are assessed on their worthiness of investigation. This results in a definitive risk selection. The Minister submits the risk reports on the basis of the definitive risk selection.'

☑ Is this clearly written down, defined and documented?
Since the court case took place, it is.

☐ Is there a regulatory body overseeing your algorithm?
The Autoriteit Persoonsgegevens (AP) should oversee this process, as well as the minister. How closely they monitor this is unclear; the court case mentioned a general monitoring process.

Autonomy

☐ Is the algorithm unable to autonomously hurt someone?
There is human interference present, however this is limited. The court did not judge on this point (6.60).

☐ Are there protocols for when humans should interact with the algorithm?
From the court case it became clear when humans should interact with the algorithm. It is not clear whether protocols are in place for this.

☒ Do citizens know how and when they can challenge a decision made by the algorithm?
According to the court case, 'it is difficult to comprehend how a data subject could be able to defend themselves against the fact that a risk report has been submitted about him or her' (6.90), mainly because the criteria are not clear, so someone cannot know or guess why the risk case was created.

Justice

Rights

☒ Are the rights and interests of all parties analysed and reflected in the algorithm?
Not much is known about the algorithm, also not how it was created and whether the stakeholders were taken into account or not. Considering the court indeed judges SyRI to have disproportional impact on private life, this is not the case .Otherwise hopefully their wishes would have been taken into account.

Fair

☐ Is the algorithm free of bias and discrimination?
The court case mentioned there is a risk of discrimination, because of the large amounts of data that may be used by the algorithm, including sensitive personal data. Whether this has been prevented enough by SyRI is unclear, according to the court.

☐ Is the algorithm free of any potential discriminating factors?
This cannot be said, since the exact criteria are unknown. There is sensitive data present in the list that may be used, like education and gender.

☐ Do team members believe the algorithm is free of any potential discriminating factors?
This question cannot be answered, since the team is unknown and cannot be discussed with.

☐ If there is any unfairness in the algorithm, is this justifiable?
According to the court case, there cannot be judged whether the unfairness in the algorithm is being dealt with sufficiently. This is because so much of SyRI is kept a secret. It does feel a

bit questionable however, mainly because SyRI was only applied to 'problematic neighbourhoods', which only increases the stigma surrounding these neighbourhoods.

Stakeholders

☐ Are stakeholders included in the development process?

There was no information found about the development process.

Accuracy

☒ How accurate is the algorithm?

The test results for SyRI were pretty disappointing. It was tested in a neighbourhood in Capelle aan de Ijssel. From here 137 risk cases were detected by the algorithm of which 41 were deemed serious by the SWZ. Of these 41, none were actually fraud[47]. This does seem like the algorithm is pretty inaccurate.

☒ What is the percentage of false positives marked by the algorithm?

In this case where the results were known, the percentage was 100%.

☒ What is the harm of the (in-)accuracy?

The harm is that innocent citizens are checked thoroughly by the algorithm. Depending on how they do this, this might harm citizens because it invades their privacy. If there are false positives, and we noted that the same people might become a false positive multiples times, this might hurt these people multiple times as well. It is important to note that there might be serious harm involved.

☑ Is the algorithm adapted to handle new scenarios?

The idea was to adapt the risk model to certain neighbourhoods. How this was done and whether it was effective is unclear and SyRI has stopped since then as well.

Effectiveness & Efficiency

☑ Can your goal be achieved by using an algorithm?

If the algorithm is tested enough and is tailored to the neighbourhoods to which it is applied, it can definitely be achieved by using an algorithm.

☑ Is this the best way to do so?

It is probably one of the more efficient ways to do it, because the algorithm is way faster than humans.

---

47 https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjb0ubFkfDrAhXJsKQKHWvxC4I4ChAWMAV6BAgGEAE&url=https%3A%2F%2Fwww.rijksoverheid.nl%2Fbinaries%2Frijksoverheid%2Fdocumenten%2Fkamerstukken%2F2019%2F12%2F20%2Fantwoorden-kamervragen-2019z21611%2Fantwoorden-kamervragen-2019z21611.pdf&usg=AOvVaw0QJa65l6WXjD2PndzLdJKL

☑ How do similar organisations (with similar solutions) perform?

The predecessor of SyRI caught €21 million euros between 2008 and 2014. This does not feel like too much. Other, less similar algorithms are present, for example in Eindhoven. Here they predict crowd movement and incidents in the inner city by using the security cameras. In Rotterdam, social media is scanned to enrichen an emergency call. Both these two projects are deemed successful; however, no precise numbers are present. It seems that very similar solutions did not work very well, but there also seems to be a potential.

☐ Are the results of the algorithm effective?

This question is a bit hard to answer, since the test results of the algorithm are unknown. In the run on Capelle aan den Ijssel, 137 risk cases were found. In a more workable situation, assuming these results would be accurate, this number could be reduced so there would be enough time to sufficiently check all these cases.

☐ Does the algorithm make efficient use of resources, in terms of money, energy, manpower and data?

The development phase of this algorithm is mainly kept a secret, so it is unknown how much money was involved here. The same goes for energy and manpower. It is unknown on what server it is located and how many people are involved with the algorithm. Whether it makes efficient use of data is highly questionable, since so much data is involved.

Explicability

Stakeholders

☑ Have you identified the stakeholders of the algorithm?

There are many people involved in SyRI. These are citizens (including customers), SZW, the minister, IB, municipalities, other organisations of whom data is included, the first and second chamber and the AP. Also, the parties who started the court case are stakeholders, albeit opposing ones. Society as a whole can also be identified, because SyRI should enhance the support base for paying taxes.

☑ Have you also identified indirect stakeholders?

These are society as a whole, citizens (who will not be checked by the algorithm) and opposing parties (i.e., those who started the court case). Because of jurisprudence, it might be that parties who use a similar algorithm are an indirect stakeholder as well, since they are dependent partly on the results of other projects as well. For example, if one algorithm turns out to be really bad, society is less likely to react positively to another algorithm.

☐ Have you identified their position?

See Table 18.

Table 18 - Stakeholders of SyRI

Citizens	Functional beneficiaries, data subjects
SZW	User of the system, developer, maintenance, data processor, data scientists
IB	Supplier of data, data processor
Minister	Political beneficiary
Municipalities & other (governmental) parties	Suppliers of data
Opposing parties	Threat agent or negative stakeholder
Society	Sponsor and indirect stakeholder
Parties with similar projects	Indirect stakeholder, political beneficiaries

From Wieringa 2014. Note that these are probably not all stakeholders, but these are all to identify with the information known.

☑ Do you know what their expectations and wishes are?

See Table 19.

Table 19 - Expectation and wishes of stakeholders of SyRI

Citizens	Society and citizens are somewhere in the middle, they would aim for catching fraudsters with reasonable amounts of data involved. Society as a whole does benefit from catching fraudsters more than individual citizens do. SyRI, however, was made to analyse all citizens, not only customers. Hence, they would want at least some form of privacy, and not have all possible data analysed and at least know what fields are analysed.
SZW	They would want to detect as much fraud as possible and probably use as much data as possible to achieve this. They are on the opposite side from the customers (who fall within the citizens group).
IB	The IB should be rather neutral: it doesn't matter how much data they need to anonymize.
Minister	The minister wants to catch fraudster to let society believe in taxes and in the system as a whole. The stakes are high for them.
Municipalities & other (governmental) parties	They want their data to be used correctly and for a good cause. They do want to see results.
Opposing parties	They simply want the algorithm to be stopped, especially in this case with the court case.
Society	Society and citizens are somewhere in the middle, they would aim for catching fraudsters with reasonable amounts of data involved. Society as a whole does benefit from catching fraudsters more than individual citizens do.
Parties with similar projects	They aim for the best use of an algorithm: as privacy friendly as possible, while catching as much fraud as possible. In any case for them the ethical and lawful side should be spotless.

☑ Is it clear how they are affected?

See Table 20.

Table 20 - How the stakeholders of SyRI are affected

Citizens	Citizens are positively affected when the algorithm works correctly. Their tax money goes to the right place. However, their data are used. They may become a false positive which may lead to extra consequences for them, such as house visits or interviews. This might also have a social effect, because people surrounding them might notice something is off and start treating them differently. It might even have an influence on economic parts, since people who are being researched might not be able to request certain benefits for example or may even receive an unfair fine.
SZW	SZW will only be affected when the project goes wrong. They must ensure that their project is legally, ethically and technically solid. If the project goes wrong, like in this case, this can lead to distrust of the society, which is very undesirable for the government. This in turn can lead to new project receiving more resistance from the people. If the project goes right, tax money can go to the right places and this might lead to new initiatives. The SZW might also gain trust from the citizens which leads to a positive image.
IB	The IB is less heavily involved. They are only affected when something goes wrong, because this might lead to mistrust.
Minister	The Minister is often seen as the end responsible and the project failing might have serious consequences for them. On the other hand, a big success can also have positive consequences, especially here, where new techniques are being applied.
Municipalities & other (governmental) parties	These parties are affected by the algorithm, because it makes their job easier. They would also save money by ending wrongful assistance. If the project goes wrong, like in Capelle aan den Ijssel, citizens in the municipality might distrust the municipality. This would be negative for the municipality again.
Opposing parties	Opposing parties will be affected when their sole purpose in to get rid of the algorithm, in this case. If they manage to, their existence would not be relevant anymore. If the algorithm would be a success and people would deem is as such, they would also be out of a purpose. Like in the court case, it ends with them being right or wrong. They will also be affected when they are more than a one-issue party, namely they can gain political benefits. If they manage to indeed oppose something, this might lead to them gaining trust from their followers or even gaining new followers. The other way around, when they fail to oppose something, this might also cost them followers or respect.
Society	Society would benefit from a properly working algorithm. They would trust the government more and their money gets used

	for good, fair purposes. If the algorithm does not work properly, it leads to distrust and they would also be less susceptive to new, similar initiatives.
Parties with similar projects	Parties with similar project are dependent on similar projects, because these can set the tone of the people. People will compare similar projects and if one goes horribly wrong, that will affect the other project as well. This new project now has to prove that they are better. The other way around, if a similar project is a huge success, this can also path the way for similar projects. Hence, they are very dependent on each other, but can also learn from each other.

☐ When will you re-evaluate this?

This cannot be answered.

Context

☑ What is the context of the algorithm?

The context lies within certain zip codes. The algorithm is applied to these zip codes. One zip code, in the Netherlands includes on average 7600 addresses[48] of citizens. The goal is to find discrepancies in the data to catch fraudsters. Before the algorithm is applied to these zip codes, it must first be allowed for by the minister. The data used for this is specified.

☑ Could this change?

This context could easily be changed, because the law allows for it. It could for example change to check only citizens who actually receive certain benefits, like social security assistance or to check other subsets of citizens. More data could also be included.

☐ When will you re-evaluate this?

This cannot be answered.

Transparency

☐ Is it discoverable for the users and data subjects how and why the algorithm made a certain decision?

The state mentioned in the court case that the creators and users of the algorithm know the decision tree, they also mentioned that it is almost impossible to recreate a decision of the algorithm, however. For the data subjects it is not desirable to have full transparency, since they might then game the system.

[48] https://www.cbs.nl/nl-nl/maatwerk/2018/49/bevolking-en-huishoudens-4-cijferige-postcode-1-1-2018

☑ Does your algorithm have full transparency (transparency design) or partially (post-hoc)?
The state does mention that SyRI is no more than a decision tree. Hence, it should be fully transparent for the users.

☐ Is it desirable to have full transparency in your algorithm in the context?
In this case, it is not desirable to have full transparency. Especially when the algorithm is a simple decision tree, because it would then be possible to game the system. For people who are marked as a potential risk it would be good to know why they were marked as such, so research can be done in an open and fair manner. If someone doesn't know of what they are accused, they cannot defend as well.

☐ Can you test the algorithm?
State mentioned that the algorithm and risk models are verified, but there is no transparency in this. In theory, the algorithm should be testable by applying it to a fake data set or an anonymous real data set. In the answers given to questions asked by the second chamber was mentioned that it is not possible to reconstruct an investigation, because all data are deleted immediately.

☐ Are there protocols for handling data and the algorithm?
This cannot be answered for sure.

☑ Is it transparent which data are used and why, and who can access them at which times?
It is to a certain degree clear which data may be used, because these are listed. However, as also mentioned by the court, there is almost no personal data that can be thought of that cannot be placed under one of the points in the list. Why these are chosen is not mentioned anywhere the author could find. Who can access the data is clear on a higher level, namely the organisations are known. Who are in the organisations or what departments within the organisations is not very clear. When people can access the data is clear. In general, this point is relatively clear.

☐ If the algorithm is malfunctioning or producing unexpected outcomes, is there an exit strategy or change protocol for adjusting the algorithm?
This cannot be said for sure.

☐ Are there protocols in case of (extreme) public outrage and who is responsible for this?
This cannot be answered.

☐ Are there other parts in or surrounding the algorithm that need protocols?
Since it is not sure which protocols are in place, it cannot be said which ones are missing.

☐ Are all protocols clear and transparent?
Protocols are not transparent; therefore, it cannot be answered whether they are clear.

Users

☐ Can the users of the algorithm explain the algorithm?
This is unclear.

☐ Are the users of the algorithm educated on how dependent and reliant they can be on the algorithm?
This is unclear.

Education

☒ Is the public educated about the algorithm and do they have a say in it?
A model information letter was created to inform citizens that SyRI would be applied to their neighbourhood. The court judged that this letter was not informative enough. Many specific parts of the algorithm are not public; the reason why the author could not answer many questions. Citizen also do not have a say in the algorithm; they do not even know what criteria there are or why and if they are considered a risk case. The state also mentioned that they would not make public the criteria, because citizens could then game the system. Citizens are of course somewhat involved, because they elect the seconds chamber who then makes a decision on the algorithm. In the end, the author would say that the public is not sufficiently educated on the algorithm and do not have a say in it, because of this.

☒ Do they know the goal of the algorithm, why you chose a certain type of algorithm, which data are used and how often the algorithm will use these data?
The goal of the algorithm is clear: 'Catching fraudsters'. There is no information provided on why this algorithm was chosen. They know about the data that may be used, if they actively look for it. It is unknown how often the algorithm will be used.

☒ Do they know what the consequences may be, who is responsible for the analysis and which quality checks are present?
It is unclear what the consequences are, as mentioned in the second chamber[47]. The public is able to know who are responsible, if they actively look for it. It is unclear which quality checks are present.

☒ Is it clear who is responsible for communication with the public?
This is unclear. There could be said that this is the same person who attended or prepared the court case. This has been anonymized and cannot be said for sure. There could also be said it is the minister, but even that would not be completely their task.

☒ Is the communication to the public done in a concise, understandable and easily accessible manner?
There was barely any communication to the public, so this was not the case. A letter was created, however, to inform citizens when their neighbourhood will be checked. This letter included limited information about data that are used, as well as which parties are included and how the feedback from the algorithm will be done. Court judged that this was not sufficient.

☒ Does your website include the fact that you perform data analysis, why you do this, what the consequences may be for citizens, whether or not you use machine learning and an explanation of this, what the legal basis is, which data sources are used, who is responsible for the analysis, what the role of third parties are in this process, which quality checks are performed, if there is human intervention in the process and which assessment frameworks are present and how they are used?
No information from the state could be found online. It might be that this information was present in the past, but has been deleted since. What could be found were reactions to questions of the second chamber, however, these were not proactively posted.

☐ Are your employees well-informed?
This cannot be answered, since no information about this could be found.

☐ Are their concerns addressed properly?
This can then also not be answered.

Citizens

☒ Are citizens informed when they are checked more extensively by the algorithm or when their data are at risk?

It became clear from the court case that citizens are not informed when they are a risk case according to the algorithm. However, a letter was created that could be sent to citizens when their neighbourhood will be checked. If citizens are a risk case, according to the algorithm, they can request this information, but it will not be proactively sent. Court judged that this was not sufficient (6.54). Added to this was that people who are not considered a risk case by the algorithm cannot know that their data were used (correctly and just).

☒ Do they have the option to raise objections to the decisions made by the algorithm?
Since they would not know that they were considered a risk case, this would be difficult.

☒ Do you remind citizens of the usage of the algorithm once in a while or at least when something changes?
This does not seem the case. Citizens are only limitedly informed when SyRI was applied to their neighbourhood.

Accountability

☑ Is it clear who created the algorithm?
The algorithm was created by the SZW. Who exactly or which department is unclear.

☐ Is it clear who is responsible for the way the algorithm works, the data that are chosen, what happens when something goes wrong, the usage of the algorithm and for creating protocols?

☐ Is it clear who is responsible for the way the algorithm works, the data that are chosen, what happens when something goes wrong and creating protocols?

This seems to be shared responsibility of SZW and the minister. However, this cannot be said for sure.

N.B. It was later found that the usage of the algorithm should be included in this question. It is not completely clear why a certain neighbourhood would be chosen to apply the algorithm to. See 6.4.1 SyRI for more information.

☑ Is it clear who is ultimately responsible for the project?
That would be the minister.

☑ Are these documented?
This is not explicitly stated, but the minister is often responsible for these types of projects.

Appendix E

This appendix applied the project framework to the case study of Gemeente Amersfoort.

Beneficence

☑ Which ethical theory fits your organisation and will be followed throughout this framework?
The municipality is often called a local government. Therefore, it makes sense that they have almost the same ethical system as the government itself. Municipalities also have the larger tasks of ordering, protecting and promoting citizens. Here also the combination between utilitarian (most happiness for most people), but also Kantian (people have duties, like working) and a rather new ethical theory: ethics of care (taking care of each other and the world). The author would say that since Gemeente Amersfoort is a rather social organisation, they lie a bit more towards the latter of the three.

☑ Is the algorithm (generally speaking) beneficial to humanity?
The algorithm of Gemeente Amersfoort has the same goal as SyRI: to stop unjust social security assistance and to make citizens trust the system. This is beneficial to humanity. An extra goal was mentioned as well, namely that an algorithm might, over time, become more objective than humans. Another benefit that was mentioned is that law enforcers might work more efficiently, because the algorithm helps them with this.

☑ Have you (re-)evaluated laws, human rights and other conventions for this?
All relevant laws have been evaluated by the municipality to see whether the algorithm would comply with them. Also, the SyRI case is closely monitored as well as the project of Gemeente Nissewaard. The municipality cannot differ from the law; however, they can differ from the ethical side related to the law. This ethical side will be evaluated here. All in all, a good look has been taken into corresponding laws and jurisprudence.

☑ What are economic, societal and environmental benefits of the algorithm?
Economic benefits are clear, a lot of money is lost due to fraud and the algorithm of Gemeente Amersfoort should be able to prevent this for this municipality. Societal benefits are that the money of the tax payer goes to the right place. This helps with citizen's trust in the government as well. Additional societal benefits are that people can be helped earlier on and will get less of a penalty when fraud is detected earlier on. Environmental benefits are unclear.

☑ Do these benefits fit your organisation and its goals?
They do. Especially since the mindset is the following. The earlier fraud is detected the earlier these people get help and assistance. Also, part of the unjustly received money must be paid back. The longer the fraud goes on the more repercussions it has therefore. Next to that, municipalities have the duty to detect and prevent fraud.

Non-Maleficenc

Privacy

☑ Which information (data) are used for the algorithm and why?

Information that are present are information about the client, their file, children, holidays, blocks, rights, exemptions, contacts, special situations, capital, debtors, requests, participation routes, activities and steps. These were chosen to remove as much of the stigmatizing factors without losing the ability to detect fraud.

☑ Is this information legal to use?

They are and they would normally, without an algorithm, also be used.

☐ Is this information ethical to use?

Most stigmatising factors have been left out from the algorithm. Gender is still present, as well as special notices (like someone getting a restraining order from the local bakery, for example). Other than that, it seems pretty logical and needed factors to use. The data is limited to these fields, which are rather clear as well. The fields that are included now, in the initial setup, are not the final fields. When the algorithm gets tested, all fields that do not add something to the decision-making process will be removed. This question should be re-evaluated after the testing.

☑ Is the amount of data used proportional to the goal?

This will be checked in a later question of this report. It is unsure whether all fields add something to the decision-making process. When looking at the data used, this does not seem unreasonable, however, since the data set is rather limited.

☑ Do all these used criteria add something to the decision-making process?

Totta Data Lab mentioned that this was indeed the case (Totta Data Lab 2020). In a later stadium, when the algorithm gets tested, this will be checked again. Criteria that do not add to the goal will be removed then. This will also be checked in a later sub question.

☐ Is the goal unreachable with less data?

The goal will probably still be reachable with less data; however, the accuracy will probably go down. This should also be checked in a later sub question. Totta Data Lab themselves mention that they only use relevant criteria in their algorithm (Totta Data Lab 2020).

☒ Does the algorithm work without training?

The algorithm does need training, since it is a learning algorithm. The algorithm will also be trained while already using the algorithm, in case it missed a fraud case.

☑ Can you use a fictional dataset or an anonymous dataset?

Both should be possible. Totta Data Lab mentioned that they use historical data, which is split into two data sets. One to train the model (70% of the data) and one to verify the model (30% of the data). After the model is trained, it will be applied to the

second data set to see how well it can detect fraud cases it has not seen before. These datasets are anonymized, because it is not needed to find the fraud cases back again.

☑ Is this dataset representative of the real world?
Historical data is real world data, so yes.

☑ Are data subjects protected from being a false positive (too often)?
From a document published by Gemeente Nissewaard (Totta Data Lab 2020), another municipality using the algorithm of Totta Data Lab, it is stated that people should be protected from being a false positive too often, because of combining multiple algorithms and by teaching the algorithm about the result of what it predicted. The latter is done by assigning a label to risk cases which mentions whether someone actually performed fraud. From this label the algorithm can then see whether it was correct and learn from it to predict better in the future.

☐ Is the impact on someone when they are a false positive acceptable?
The impact on someone in rather large. This is because quite a lot is possible then, like house visits and interviews. If someone is considered a risk case, Gemeente Amersfoort will start with a desk research first. Possible false positives may then be filtered out already. Then they can have a talk with them, so they know they are considered a risk case. There was stated during the interviews that most signals come in from people close to the customer, and these signals would sometimes also come in because people were in a fight with each other, for example. The impact on people when using the algorithm versus without the algorithm is quite similar. However, it is still something to keep in mind.

☐ Is the algorithm also privacy friendly for people with background knowledge or when combined with other data sources?
All identifiable data are pseudonymized. With the correct background information, these information are still identifiable, especially in unusual circumstances (people with many children, for example). However, this information (combined with background information) should only accessible for people who would otherwise also know this information. Also, only a limited number of people within Amersfoort can access the data, which they can also access without an algorithm. The data at Totta Data Lab is pseudonymized. People working here could in theory be able to recognize exceptional cases within the data. However, these people are trusted and a processing agreement has been signed as well[49].

N.B. please note that there was found that this question should be altered. With this alteration, this question would be checked/green. See 6.4.3 Amersfoort for more information.

[49] https://www.Totta Data Labdatalab.nl/portfolio-item/klantcase-wil/

☑ Will the algorithm still be privacy friendly in 20 years?
Since the datasets are not public, yes it should be the case.

☑ Is there checked whether there are people in the team who can imagine a future scenario in which the results of your project can be misused?
There were discussions about the ethical and legal side of this project. These have been resolved. These were more focused on the privacy aspect of the data. Within the team, regular updates on this are sent around to keep everyone on point, especially with the current media attention for the same algorithm in Gemeente Nissewaard. The criteria are also re-evaluated to check whether they are still okay to use, since this may change over time. Also, regular meetings with the team and with Totta Data Lab were held during the project.

☑ Have you conducted a DPIA?
Yes, this has been conducted.

☑ Have you discussed your ethical ideas and concerns with a privacy officer or an independent organisation?
The ethical aspects are discussed within the team of Gemeente Amersfoort, with the author of this report, with Totta Data Lab and with the privacy officer of Gemeente Amersfoort. It is arguable whether the latter would be independent, however, they would not personally gain anything with the algorithm or with catching more fraudsters, so the author would argue that this has been done.

Misuse

☐ Do you have clear protocols regarding what the algorithm may and may not do and how humans should interact with this?
There are protocols regarding who have access to the data (and the algorithm) and where the data goes and flows, there are also protocols on how the law enforcer and functional management should interact with data and the fraud detection process. These protocols, however, should be updated once the algorithm is further developed.

Future

☐ Do you have clear time paths for re-evaluating this framework?
As of now, this is unclear, since Gemeente Amersfoort doesn't know whether it will actually continue with the development of the algorithm and since this framework is filled in by the author, and not by the municipality. There was mentioned that the ethical side should be re-evaluated regularly, since the opinion of the public may change over time.

Security

☐ Is the data secure for the outside and for the inside?
This question cannot be answered completely, since the author does not know the full details. The data must be sent between Totta Data Lab and Gemeente Amersfoort. With another municipality (Gemeente Nissewaard), this is done via a secured SFTP connection. This will be resolved in the same way for Gemeente Amersfoort. There are people working on the security for both Gemeente Amersfoort and Totta Data Lab. Although the question

arises whether something can be secure at all, the author believes this seems plausible at least. For the inside protocols need to be made. This should be done before fully implementing the algorithm. The functional management alters the data in such wat that there should be little to no possibility of misuse present anymore for the law enforcers. With the protocols also in place, there seems to be reasonable security. The protocols must be updated once the algorithm is being tested.

Up to date

☑ Will the algorithm be updated regularly?

Approximately every quarter of a year the data will be handed from Gemeente Amersfoort to Totta Data Lab and vice versa. With these data, the algorithm can be updated.

☑ Do you know what 'good behaviour' entails for your algorithm?

From earlier studies between Totta Data Lab and the municipality of Gemeente Nissewaard, it was found that an accuracy of 50-60% of the top 10 fraud risk cases is considered 'good' (Totta Data Lab 2020). Furthermore, it is mentioned here that the same cases should not come forward time after time. Further technical documentation on the data and on the algorithm, itself are found in this document as well. Every organisation to work with this algorithm of Totta Data Lab also gets this documentation. It seems pretty known what good behaviour entails.

☑ How can you check this?

The data is analysed by the same party as who created the algorithm in the first place. They receive real world feedback from the predictions of the algorithm. When the accuracy goes down too much, this is an indicator that either all fraudsters have been caught or that there is something wrong with the algorithm. Furthermore, if false positives come up by the algorithm over and over again, this also indicates that there is something wrong. Some focus is also given to so-called easter-eggs in the data; criteria that already mention that there is a fraud case. If the accuracy becomes 100%, this stands out.

☑ Are the data (actively) being kept up to date?

This data is from the municipality. Both they and the citizens have the duty to keep this up to date.

☑ Are the data of good quality and well organised?

Only data fields are used of which at least 50% of the data is filled in. Some data forms are changed to make it more understandable for the algorithm. For example, 'Has children?' can be answered either with a 1 (yes) or a 0 (no), instead of text.

☐ Are all protocols up to date?

Not yet, some protocols need to be altered still. Like the one of the law enforcers and that of the functional management, this can be done when the algorithm gets tested.

Developing

☑ Is the algorithm developed by skilful, rational and responsible people?

The people at Totta Data Lab are definitely skilful; they make other IT solutions as well and have all finished interesting studies. From the interview with Jesse earlier in this report, they seem to have thought about their solution a good bit, they also mentioned the flaws of their algorithm for example. They also took the time to discuss their algorithm at Gemeente Amersfoort. The author would say so, for this question.

☑ Do you also know the weaknesses of your team?
When asked, it was mentioned that the main weakness is that they do not develop software of algorithms themselves and are therefore always dependent on another party. There is no way of solving this. This should thus be kept in mind.

☐ Is the algorithm tested extensively?
The algorithm is currently (publicly) applied to three organisations: Gemeente Nissewaard, Orionis and Lekstroom. Gemeente Nissewaard is working on an official investigation on the algorithm. The results of this are of course interesting for Gemeente Amersfoort as well. What the results of these 'tests' or actually real-world applications are is nowhere to be found. The three organisations are still working with Totta Data Lab, which is a good thing. The algorithm itself is also tested on the historical data. The algorithm is thus not yet tested extensively, but it is tested as much as possible currently.

☑ Is the development documented properly?
Yes, all organisations also receive this documentation.

Access
☑ Is it clear who can access the data and when?
The functional managements send pseudonymized data to Totta Data Lab who then uses the data to apply the algorithm to. Then the processed data comes back in from Totta Data Lab at the functional management, who sees the top 10 and decrypts the data. They also choose 3 customers at random to send through to the law enforcers. The law enforcer thus receives 13 customers to research. Furthermore, the law enforcers at Gemeente Amersfoort are able to see the data, when they research someone. They do not see what the algorithm stated about them (risk percentage), they only receive 13 customers to research.

☑ Is this clearly written down, defined and documented?
A data flow diagram has been created to create a clear overview.

☑ Is there a regulatory body overseeing your algorithm?
The algorithm of Totta Data Lab was sent to at least The Ministry of the Interior and Kingdom Relations, Ministry of Justice and Security and the Ministry of Social Affairs and Employment[50]. Federatie Nederlandse Vakbeweging (FNV) noticed correctly that these parties also watched SyRI. However, just like with SyRI, the author would argue that there is

[50] https://www.binnenlandsbestuur.nl/sociaal/nieuws/rookgordijn-rondom-fraudesysteem-Gemeente Nissewaard.13026523.lynkx?tid=TIDP329789X862691AA25804059A3753F14DB95B20EYI5&utm_campaign=BB_NB_Dagelijks&utm_medium=email&utm_source=SMG&utm_content=854_24-04-2020

a regulatory body overseeing the algorithm, however, municipalities also have a large responsibility themselves and should stay critical.

Autonomy

✅ Is the algorithm unable to autonomously hurt someone?
The algorithm gives a top 10 of people who have the higher chance to fraud. These 10 are then researched by a human. This research does have impact. Normally people can also be randomly selected to be researched, which has similar impact. People can also be 'selected' by someone calling to give a signal about them, which might be real or not. Hence, the algorithm can indirectly hurt someone by predicting they might be a fraudster; however, a human will always follow up on this. The algorithm should not hurt someone more than they would be without an algorithm. The percentage of false positives goes down with the use of an algorithm[51], so it might be that people are less hurt with the algorithm compared to the current situation. Therefore, the algorithm should not be able to autonomously hurt someone.

☐ Are there protocols for when humans should interact with the algorithm?
There are protocols in place already for the functional management and the law enforcers, however, these should be updated when the development of the algorithm is in a later stage. The data flow diagram covers a part of this question already, however.

✅ Do citizens know how and when they can challenge a decision made by the algorithm?
Citizens are informed when they are considered a risk case. Also, a conversation is then planned with them to discuss this. Even if citizens may not know beforehand how to do this, this will be resolved for them by having this conversation.

Justice

Rights

✅ Are the rights and interests of all parties analysed and reflected in the algorithm?
While tailoring and researching the algorithm, Gemeente Amersfoort discussed it with some parties. For example, with the client council, who represents the customers. As well as the alderman, local council (*nl: gemeenteraad*), this research, a research of another municipality and people within the team of Gemeente Amersfoort. Therefore, the rights and interests of most, if not all, parties seem taken into account.

Fair

☐ Is the algorithm free of bias and discrimination?
At the start it will not be free of bias and discrimination, because it learns from the results made by humans, who are not free of bias and discrimination. However, the longer the algorithm will be used, the more objective it becomes, because it will learn more patterns. By using Random Forest, the algorithm will randomly create decision trees, which makes the possibility of bias less. The algorithm itself should at least not be more discriminating than

[51] https://www.sociaalweb.nl/cms/files/2018-12/machine-learningv1-002-.pdf

humans. Thus, the algorithm is not free of bias and discrimination, but it will, when first implemented, be as biased and discriminating as humans. The algorithm will be applied to customers equally. This contrary to SyRI, which was applied to 'bad' neighbourhoods only. This point can become green, or checked, when the algorithm is in place for a good amount of time already.

☒ Is the algorithm free of any potential discriminating factors?
There are at least gender and special situation that can be remarked as potentially discriminating. Whether they will stay in the final algorithm is not sure. The municipality is allowed to use these data, so if they really add something to the decision-making process, they will probably stay in. Hence, for now this question will be marked as red, but this may change when this framework is re-evaluated.

☒ Do team members believe the algorithm is free of any potential discriminating factors?

The two mentioned above were also mentioned by the team.

✅ If there is any unfairness in the algorithm, is this justifiable?
Especially in the beginning there will be unfairness in the algorithm, since the algorithm is trained on human, historical data. Most unfairness should be trained out over time. The author believes this to be justifiable, because it is not more biased or discriminating than without an algorithm.

Stakeholders

✅ Are stakeholders included in the development process?
To a certain extent they are. This is done via the client council, local council, alderman and via the municipality itself. Customers do not directly have a say in it, but are represented by the client council.

Accuracy

✅ How accurate is the algorithm?
The accuracy mentioned by Totta Data Lab for Gemeente Nissewaard was 50-60% (Totta Data Lab 2020). During another presentation, an accuracy of 20% was given[51]. It was mentioned by Gemeente Amersfoort that they expect about 20% of the signals of the algorithm to be indeed fraud, based on a business case from Totta Data Lab for Gemeente Amersfoort. This is more accurate than without an algorithm.

✅ What is the percentage of false positives marked by the algorithm?

At Gemeente Nissewaard, of the 29 cases marked by the algorithm, 15 were not further investigated. So, there could be said that 50% is a false positive. At Gemeente Amersfoort, there should be at least 3 false positives every time, since these are chosen at random.

☐ Is there no serious harm involved in the (in-)accuracy?

It is not clear how quick the cases at Gemeente Nissewaard mentioned above

were stated not to be fraud, i.e., whether they were thoroughly researched or immediately unmarked as a possible fraud case. There will always be a bit of an inaccuracy, since there will also be 3 random cases in the possible fraud pile. The harm of the inaccuracy should not be bigger than without the use of an algorithm. With being a false positive, the algorithm should also be trained, such that it can learn from the incorrect prediction.

☑ Is the algorithm adapted to handle new scenarios?
Yes, this should be the case, since the algorithm is trained whenever new cases of fraud occur. Of course, those checking on fraud are always one step behind the ones actually performing fraud, but one case caught can lead to the algorithm learning new patterns.

Effectiveness & Efficiency
☑ Can your goal be achieved by using an algorithm?
Yes, that seems possible.

☑ Is this the best way to do so?
It is probably one of the more efficient ways to do it, because the algorithm is way faster than humans. It the end it should also become more objective than humans.

☑ How do similar organisations (with similar solutions) perform?
The preliminary results of Gemeente Nissewaard were published[51].
Furthermore, they will start researching the algorithm with an independent research organisation. From this Gemeente Amersfoort can also learn. Their performance in the first quartile of using it was 5 fraud cases out of 10 marked by the algorithm were actually fraud.

☑ Are the results of the algorithm effective?
They are, because all 13 cases can be researched. There should be, in theory, always at least 3 that are not fraud. This makes it feel inefficient, but it also keeps the law enforcers critical and objective. With the algorithm, the pond in which they fish (check) fraudsters should be more specified with an algorithm. It seems pretty effective. Without the 3 random cases, it could be more effective, but this would lose on objectivity and sharpness. It seems logical to make that trade-off in favour of the objectiveness.

☑ Does the algorithm make efficient use of resources, in terms of money, energy, manpower and data?
In terms of money, the algorithm has to discover only 1 fraud case in order to earn its money back. For energy, it is more power consuming than without an algorithm, but the algorithm is applied to the data on a server of Totta Data Lab. The more algorithms they make for different organisations, the more efficient that goes. Whether their server is a physical one or not is unclear. If it is a physical one and considering the number of projects they have, it would be reasonably efficient. If it is in the cloud, it would be more energy efficient. Considering the current time being, it is more assumable that their server is in the cloud, but this cannot be said for sure. The manpower needed is a bit more than without an algorithm, since the functional management needs to perform extra steps, but it should also increase effectiveness of detecting fraudsters. Last in terms of data, it is approximately equally efficient with or without an algorithm. Law enforcers are able to see more data than the

algorithm, but in the total process the same data are used, only a small subset is used by the algorithm.

Explicability

Stakeholders

☑ Have you identified the stakeholders of the algorithm?

Involved with this project are 15 stakeholders. Who they are and what they do is listed below. Stakeholder identification was also done by Gemeente Amersfoort to know who to the distribute what (dept of) knowledge to about this project. For example, the functional management must know more than the customers about the technical details.

Project manager: Is responsible for the planning and successful realisation of the project. They also are the direct contact person from Gemeente Amersfoort for this thesis.

Information provision advisor (*nl: IV adviseur)*: Ensures the right data and applications are used. They look at the project from the data point of view.

Data scientist: Looks into the algorithm of Gemeente Amersfoort.

Supplier: This is Totta Data Lab in this case, providing the algorithm as a solution for fraud detection.

Legal advisor: Works within Gemeente Amersfoort to make sure everything is legally allowed.

Communications advisor: Give advice regarding the communication from the Alderman (*nl: Wethouder)*.

Enforcers: Research potential fraudsters and track them down. There are six enforcers working with Gemeente Amersfoort.

Citizen managers: They work at the municipality and process the applications for social assistance and check their justice. They are also the first point of contact for the customer (i.e., people who receive social security assistance).

Functional management: Are the application managers; they have access to the dataset and provide the enforcers with information.

Team manager: Manages the team, is also involved with this thesis.

Department manager: Manages the department Job, Income and Care (*nl: Werk, Inkomen en Zorg)* in this case. The team manager falls within this department as well. They gave the order to start this project.

Alderman: Is the public director within a municipality. They are comparable to ministers within the Cabinet, they are connected to a political party and are connected to a certain topic. They decide upon continuing this project or not.

Citizens of Gemeente Amersfoort: Their data will be checked via the algorithm eventually, so they are involved with this project. People who do get social security assistance are called customers.

Client council: In this council are customers of Gemeente Amersfoort. These can be organisations like the Federatie Nederlandse Vakbeweging[52] (FNV), social assistance customers and representatives.

Municipality council: Consists of democratically chosen representatives of the municipality

Society: They benefit when as many fraudsters are caught, because the tax system then works in the right way. They also gain trust in the government when fraud is stopped.

Opposing parties: There are some opposing parties who actively try to stop these types of algorithms.

From 2.5 Stakeholders.

☑ Have you also identified indirect stakeholders?

Yes, these are the society, opposing parties and parties with similar projects. Opposing parties in this case can be found in the FNV[53] and bijvoorbaatverdacht.nl[54]. Parties with similar projects are, for example, SyRI, Gemeente Nissewaard, Lekstroom and Orionis.

☑ Have you identified their position?

See Table 21

Table 21 – Position of stakeholders of Gemeente Amersfoort's algorithm

Project Manager	Consultant
Information provision advisor	Consultant
Data scientist	Consultant
Supplier (Totta Data Lab)	data processor, supplier of data, maintenance, data processor
Legal advisor	Consultant
Communications advisor	Consultant
Enforcers	Data processor
Citizen managers	Supplier of data
Functional management	Supplier of data, data processor
Team manager	Consultant
Alderman	Political beneficiary
Department manager	Client

[52] https://www.fnv.nl/

[53] https://www.fnv.nl/nieuwsbericht/sectornieuws/uitkeringsgerechtigden/2020/06/fnv-sceptisch-Gemeente Nissewaard-belooft-burger-transparan

[54] https://bijvoorbaatverdacht.nl/Gemeente Nissewaard-weigert-alsnog-beloftes-transparantie-na-te-komen/

Citizens of Gemeente Amersfoort	Functional beneficiaries, data subjects
Client council	Consultant
Municipality council (local council)	Political beneficiary
Opposing parties	Threat agent or negative stakeholder
Society	Sponsor and indirect stakeholder
Parties with similar projects	Indirect stakeholder, political beneficiaries

From (Wieringa 2014)

Note that there are more consultants listed for this algorithm than for SyRI. This is mainly because more information is present in this case and it is thus possible to go more into detail here.

☑ Do you know what their expectations and wishes are?

See Table 22.

Table 22 – Wishes of stakeholders of Gemeente Amersfoort's algorithm

Project Manager	They mainly want the project to as smoothly as possible. The project should also go correctly, however, and there must be consulted with the right people.
Information provision advisor	They want the right data to be used in the right manner. They have to discuss this with the supplier, Totta Data Lab and with other important parties at the municipality.
Data scientist	The data scientist looks at the algorithm and can advise from that perspective.
Supplier (Totta Data Lab)	Totta Data Lab want the algorithm to perform as good as possible. Ideally, they would want an accuracy of 100% (not considering the 3 random cases) and to discover new types of fraud. They mainly want to prove the algorithm works such that other municipalities would want the algorithm as well. They do not necessarily want to use as much data as possible, since they must also make sure that what they do is ethically good.
Legal advisor	The legal advisor wants the algorithm to be legally sound. They will compare it to the laws and regulation and to other, similar cases to give an advice on the algorithm.
Communications advisor	They give advice on the communications of the alderman. Mainly the reputation of the alderman in of their concern.
Enforcers	They want to catch as many fraudsters. However, another very important goal of them is to do this in a very customer friendly way, with a lot of communication between them and the customer. They do want the system to be fair.
Citizen managers	They can give a signal to the law enforcers when they suspect fraud from a client/customer. In the end they would not care too much about the algorithm, but they do want the system to be fair as well.
Functional management	They have to prepare the data for sending in to the supplier and to the law enforcers. They would not care too much

	about the algorithm, but they do want the data to be correct and secure.
Team manager	The team manager is in line with the law enforcers.
Alderman	They want as many fraudsters to be stopped as possible to let people retain trust in the government. This should be done in an ethical way such that not too many data are used and to prevent public outrage.
Department manager	They want as many fraudsters to be stopped as possible to let people retain trust in the government. This should be done in an ethical way such that not too many data are used and to prevent public outrage.
Citizens of Gemeente Amersfoort	Society and citizens are somewhere in the middle, they would aim for catching fraudsters with reasonable amounts of data involved. Society as a whole does benefit from catching fraudsters more than individual citizens do. Whereas citizens are more concerned about their (individual) privacy.
Client council	They represent the customers of Gemeente Amersfoort. Hence, they should be approximately in line with them.
Municipality council (local council)	They want to be somewhat in line with their party's ideologies. This depends a bit on how progressive the party is. Algorithms have been negatively in the news, due to the Belastingdienst[55] and SyRI.
Opposing parties	They simply want the algorithm to be stopped, by going to the media with critiques.
Society	Society and citizens are somewhere in the middle, they would aim for catching fraudsters with reasonable amounts of data involved. Society as a whole does benefit from catching fraudsters more than individual citizens do. Whereas citizens are more concerned about their (individual) privacy.
Parties with similar projects	They aim for the best use of an algorithm: as privacy friendly as possible, while catching as much fraud as possible. In any case for them the ethical and lawful side should be spotless.

☑ Is it clear how they are affected?

See Table 23.

Table 23 - How the stakeholders of Gemeente Amersfoort's algorithm are affected

Project Manager	The project manager will gain reputation internally when the project goes right and vice versa.

[55] https://www.ad.nl/politiek/affaire-kinderopvangtoeslag-bereikt-hoogtepunt-zes-vragen-over-wat-er-aan-de-hand-is~a0975385/

Information provision advisor	They are not closely involved with the algorithm. If their advice is very solid, however, they might gain some internal reputation (and vice versa).
Data scientist	They are not closely involved with the algorithm. If their advice is very solid, however, they might gain some internal reputation (and vice versa)
Supplier (Totta Data Lab)	Totta Data Lab will gain reputation (and possibly new customers) when the algorithm performs well. The other way around, they could lose customers and reputation. This is of direct impact on their revenues.
Legal advisor	They are not closely involved with the algorithm. If their advice is very solid, however, they might gain some internal reputation (and vice versa)
Communications advisor	They are not closely involved with the algorithm. If their advice is very solid, however, they might gain some internal reputation (and vice versa)
Enforcers	They will receive an extra 13 cases from the algorithm. More than that should not change. The algorithm may, over time, make their job easier if less fraudsters are present. This might also make their job more positive, as they mentioned themselves, they would like to give more attention to the customer, not only when they are a suspect.
Citizen managers	There should not change anything for them.
Functional management	They will receive more work, since they have to alter the data to give it to Totta Data Lab and to the law enforcers.
Team manager	If the project goes well, they would gain trust from within the organisation. For the team manager, this is a bit more important than for the other stakeholders, since this entails a larger part of their job.
Alderman	The alderman would often been seen as the end responsible for the project, together with the department manager. When the project goes right, they gain reputation, when it goes wrong, they lose it. Hence, they will be cautious, but interested.
Department manager	The alderman would often been seen as the end responsible for the project, together with the department manager. When the project goes right, they gain reputation, when it goes wrong, they lose it. Hence, they will be cautious, but interested.
Citizens of Gemeente Amersfoort	Citizens are positively affected when the algorithm works correctly. Their tax money goes to the right place. However, their data are used. They may become a false positive which may lead to extra consequences for them, such as house visits or interviews. This might also have a social effect, because people surrounding them might notice something is off and start treating them differently. It might even have an influence on economic parts, since people who are being

	researched might not be able to request certain benefits for example or may even receive an unfair fine.
Client council	The client council might gain some trust from the public when the project goes well, and vice versa.
Municipality council (local council)	They might gain trust from the public when the project goes well or vice versa. For them, this is a bit more important than for most other stakeholders, since they are elected every few years. Hence, if the public thinks they have failed, they might vote on another party. This can be very bad for them, since they cannot have influence on the municipality like they used to. Therefore, big changes, like this project, will be encountered cautiously.
Opposing parties	Opposing parties will be affected when their sole purpose in to get rid of the algorithm, in this case. If they manage to, their existence would not be relevant anymore. If the algorithm would be a success and people would deem is as such, they would also be out of a purpose. Like in the court case, it ends with them being right or wrong. They will also be affected when they are more than a one-issue party, namely they can gain political benefits. If they manage to indeed oppose something, this might lead to them gaining trust from their followers or even gaining new followers. The other way around, when they fail to oppose something, this might also cost them followers or respect.
Society	Society would benefit from a properly working algorithm. They would trust the government more and their money gets used for good, fair purposes. If the algorithm does not work properly, it leads to distrust and they would also be less susceptive to new, similar initiatives.
Parties with similar projects	Parties with similar project are dependent on similar projects, because these can set the tone of the people. People will compare similar projects and if one goes horribly wrong, that will affect the other project as well. This new project now has to prove that they are better. The other way around, if a similar project is a huge success, this can also path the way for similar projects. Hence, they are very dependent on each other, but can also learn from each other.

☐ When will you re-evaluate this?

This is unclear, but Gemeente Amersfoort indicated that they realise ethics may change over time and that they should check this regularly. Exact data cannot be given for this.

Context

☑ What is the context of the algorithm?

The context of the algorithm is to apply it to customers of social security assistance of Gemeente Amersfoort. The goal is to find discrepancies in the data to detect fraud. The latter is done with a preselection of the algorithm and a further investigation of the law enforcers of Gemeente Amersfoort.

☐ Is this unable to change?
This context could change if this were to applied to other people or other types of discrepancies (think about permit fraud for example) or when more data fields were to be added. However, what is and is not allowed with this algorithm is well defined, which makes it less easy to change the context.

☐ When will you re-evaluate this?
This is unclear, this this framework was not filled in by Gemeente Amersfoort. Gemeente Amersfoort indicated that they realise ethics may change over time and that they should check this regularly. Exact data cannot be given for this.

Transparency

☐ Is it discoverable for the users and data subjects how and why the algorithm made a certain decision?
Totta Data Lab worked on a way of making the algorithm more transparent. The results are now both reproducible and traceable, by means of explainable AI. This is not discoverable by the law enforcers, since they need to be objective. It is discoverable by Totta Data Lab, however. Customers (data subjects) can probably request their data at the municipality, this should be researched however, to ensure this. They do not automatically get to see why they are marked as a fraud case. If this were the case, this could also influence the law enforcer.

☑ Does your algorithm have full transparency (transparency design) or partially (post-hoc)?
There is post-hoc transparency involved, mainly because the algorithm randomly creates a decision tree.

☑ Is it desirable to have full transparency in your algorithm in the context?
No, it is not, because users can then game the system. There should be transparency to some extent, so per case to the users for example. This is so users can still see why they were considered to be a risk case, without the circumvention possibility.

☑ Can you test the algorithm?
This has been done with the training dataset and this is continuously done by learning the algorithm about fraud cases it missed.

☐ Are there protocols for handling data and the algorithm?
There are protocols for who can see the data and when. Protocols still need to be updated for the law enforcer and functional management specifically. The technical description included with the algorithm explains about what the algorithm can and cannot do and with which data this may happen.

☑ Is it transparent which data are used and why, and who can access them at which times?

Yes, see above.

☑ If the algorithm is malfunctioning or producing unexpected outcomes, is there an exit strategy or change protocol for adjusting the algorithm?

This lies mainly in the hands of Totta Data Lab. The algorithm will first be tried for a year at Gemeente Amersfoort. After this year, the algorithm will be evaluated. This evaluation is seen as the protocol for handling unpredictable situations of the algorithm.

☒ Are there protocols in case of (extreme) public outrage and who is responsible for this?

There are not. There was indicated that these might have to be created. Especially after the reaction on the algorithm of Gemeente Nissewaard, where customers of that municipality are asked to send a letter to the municipality to request their data. As well as other negative reactions on the algorithm. It is good to prepare for that.

☒ Are there no other parts in or surrounding the algorithm that need protocols?

As mentioned, a protocol in case of extreme public outrage might be useful. It might also be good to create a protocol to document how and to what extent people can request their data (for example risk score and indicators).

☑ Are all protocols clear and transparent?

They seem to be, some include also infographics to make them more easily understandable.

Users

☑ Can the users of the algorithm explain the algorithm?

When the interviews were held all parties seemed pretty updated on what the algorithm would do. The exact algorithm has not been created yet, but the steps until this point seemed clear for all parties spoken with. A communication plan has been created to keep track of who should be able to explain what. The functional management should for example be able to know about the algorithm more in dept than the law enforcer.

☑ Are the users of the algorithm educated on how dependent and reliant they can be on the algorithm?

Yes, they are, since they created the algorithm. For the law enforcers at Gemeente Amersfoort it should be the case as well. The risk cases they get from the algorithm are not only the ones the algorithm predicted, but also 3 random cases. The risk indicator is also removed, so they do not know who was the highest risk case. They thus should know that they cannot be super reliant on the algorithm, since there should also be three non-fraud cases. This should keep them as objective as possible. Protocols will also be updated accordingly.

Education

☐ Is the public sufficiently educated about the algorithm and do they have a say in it?

Currently they do not know about the development of the algorithm. They do have a say in it via the client council, who represent the customers of Gemeente Amersfoort. Once they are informed, the information is public. It is currently not yet public, because it is not entirely sure whether the project will be continued with and not all details are final. Informing the public now can lead to unnecessary outrage about, for example, a data field which would be removed in a later version. However, there is something to say about informing the public about looking at such an algorithm. This question should be re-evaluated in a later stadium, to see whether this has actually happened.

☐ Do they know the goal of the algorithm, why you chose a certain type of algorithm, which data are used and how often the algorithm will use these data?

As of now they do not know, since it is not sure whether the algorithm will indeed be implemented. When the algorithm will be implemented, the technical documentation will be published. This includes the data used, the algorithm chosen and why, the goal and how often everything is checked. This question should also be re-evaluated in a late stadium.

☐ Do they know what the consequences may be, who is responsible for the analysis and which quality checks are present?

As of now not yet, however the technical documentation includes the consequences, who is responsible and to a certain extent the quality check. The latest could be more explicit. It was mentioned that Gemeente Amersfoort wanted to include a text to explain the technical documentation a bit more. These quality checks are something that could be included here (i.e., 'How do we ensure the quality of our data and algorithm'). This question should also be re-evaluated later on.

☑ Is it clear who is responsible for communication with the public?

There are multiple people within the project team responsible for the communication with the public. Amongst them are the alderman, communication advisor and the spokesman of the alderman. All questions from the council will come in at one person.

☐ Is the communication to the public done in a concise, understandable and easily accessible manner?

The information will be mainly shared on the website of Gemeente Amersfoort. Other communication will refer to the website for more information. With other communication there can be thought of interviews, extra information in the rules booklet customers receive when requesting socials security assistance, a council information letter, letters sent by the client council and via a special portal for people with minimum income. The author would therefore argue that it is easily accessible. Extra thought is put on the understandable part by including extra information with the technical documentation that will be published. Whether these are concise should be evaluated when they are created.

☐ Does your website include the fact that you perform data analysis, why you do this, what the consequences may be for citizens, whether or not you use machine learning and an

explanation of this, what the legal basis is, which data sources are used, who is responsible for the analysis, what the role of third parties are in this process, which quality checks are performed, if there is human intervention in the process and which assessment frameworks are present and how they are used?
The technical documentation will be published on the website of Gemeente Amersfoort. This document includes the fact that data analysis is performed, whether or not machine learning is used, which data sources are used, who is responsible for the analysis, what the role of third parties are, (a bit on) what quality checks are present and if there is human intervention. Information on false positives and false negatives is available, but no information about the impact of being either is present in this document. It is documented why certain algorithms are chosen and their research is documented properly, so for the assessment frameworks it seems okay. Why data analysis is performed, what the consequences are for citizens and the legal basis and an extension of the explanation of the quality checks should be included in the accompanied letter.

☑ Are your employees well-informed?
There have been multiple meetings with employees about the algorithm. Relevant employees are informed and other employees will be informed later on. A communication plan has been created for this.

☑ Are their concerns addressed properly?
There were some concerns which are regularly discussed. Updates about, for example, Gemeente Nissewaard are also shared when something relevant passes by.

Citizens

☑ Are citizens informed when they are checked more extensively by the algorithm or when their data are at risk?
Citizens are informed when they are considered a risk case by having a conversation with them. Legally, citizens are informed when their data are at risk and this is also done in this case.

☑ Do they have the option to raise objections to the decisions made by the algorithm?
They are able to do this during the conversation or later in the research period. They might also be researched if the algorithm does not think they are a risk case, because they are a random case, or because of another signal for example.

☒ Do you remind citizens of the usage of the algorithm once in a while or at least when something changes?
Customers are not informed periodically. Depending on the change they are informed about this. This mainly goes for larger changes and may or may not be directly.

Accountability

☑ Is it clear who created the algorithm?
This is clear, since this is Totta Data Lab. This is also described in the technical documentation.

✅ Is it clear who is responsible for the way the algorithm works, the data that are chosen, what happens when something goes wrong and for creating protocols?
It is clear who is responsible for the way the algorithm works and for the data that are chosen. Protocols are created by both Gemeente Amersfoort and by Totta Data Lab, depending on the protocol. Protocols for when something goes wrong are in place for both the inside (when to report something and to whom) and the outside (via the processing agreement).

N.B. It was later found that the usage of the algorithm should be added to this question. For Gemeente Amersfoort the usage of the algorithm is clearly defined by both Totta Data Lab and Gemeente Amersfoort, namely once every three months within the context described earlier. See 6.4.1 SyRI for more information.

✅ Is it clear who is ultimately responsible for the project?
The alderman is eventually the one saying yes or no to the project. The department manager is the person requesting this project in the first place. They together would be responsible for the project itself. Other people would of course be responsible for sub parts of the project, such as making sure the project goes well (project manager) or advising on the legal aspects (legal advisor).

✅ Are these documented?
This has more to do with the structure of the municipality itself. The general structure is described by multiple websites.

Appendix F

This appendix describes which attributes of which tables were used for the creation of the algorithm of this research. Marked attributes were used.

Requests	**Activities**	**Special situation**	**Blocks**	**Client**
Client number	Client number	Client number	Client number	Client number
Code regulation	Route number	Start date	Dossier number	Date registration
Code group	Municipality	End date	Start date block	Indication sex
Description	Activity number	Special situation code	End date block	Birthdate
Date application	Activity	Description	Reason code	Date of death
Date settlement	Activity description		Reason description	Indication marital status
Code phase	Start activity			Marital status
Date time settlement	End activity			
Code handling	Activity status			
Decision id	Dossier number			
Decision				

Contact moments	Debtors	Dossiers	Icosig
Client number	Client number	Client number	Client number
Date registration	Debtor number	Dossier number	Code regulation
Date handling	Date registration	Client number partner	Code group
Code action	Code group	Date registration	Description
Number reply	Type claim	Start date	Request date
	Code Subcategory	End Date	Settlement Date
	Start Date Claim	Code Regulation	Code Phase
	End Date Claim	Regulation Circumstances	Date Time Handling
	Status Claim	Code Group	Handling Code
	Status Claim Partner	Group Circumstances	Decision Id
	Budget Balance	Indication Marital Status	Decision
	Decision Number	Code Handling	
	Start Date	Code Handling Partner	
	End Date	Code Reason End	
	Indication Cessie	Code Reason End Partner	
	Dossier number	Indication Cohabitation Conditions	
		Cohabitation Conditions	
		Indication Housing	
		Housing	

Children	Participation process	Rpo	Steps
Client number	Client number	Client number	Client number
Child number	Process number	Start date	Date steps
Reference number	Start process	End date	Code steps
Birthdate	Kind of process	Indication rpo	Step description
Date of death	Description kind of process	Code rpo	
Indication resident	Type of process	Reason description	
Indication living situation	Start process type		
	End process type		
	Code ending		
	Date step		
	Code step		
	Step description		
	Dossier number		

Vacation	Capital
Client number	Client number
Start date	Reference date
End date	Budget balance
Indication person	Code balance
	Indication owner

Appendix G

This appendix describes which attributes were removed after attribute selection.

* = altered with the feature selection of removing < 7

** = altered with the correlation selection

Grey marked = removed with the correlation selection

Requests	**Activities**	**Special situation**	**Blocks**	**Client**
Client number	Client number	Client number	Client number	Client number
Code group*	Activity*	Start date	Start date block	Date registration
Date application**	Start activity**	End date	Reason code	Indication sex
	End activity	Special situation code*		Birthdate
	Activity status*			Indication marital status

Contact moments	**Debtors**	**Dossiers**	**Icosig**
Client number	Client number	Client number	Client number
Date registration	Date registration	Dossier number	Request date
Date handling	Code group*	Code Regulation	Decision Id
Code action*	Status Claim	Code Group*	
Number reply*		Indication Cohabitation Conditions	
		Indication Housing	

Children	**Participation process**	**Rpo**	**Steps**
Client number	Client number	Client number	Client number
Birthdate**	Start process	Start date	Date steps
	Kind of process*	End date	Code steps*
		Code rpo*	

Vacation	**Capital**
Client number	Client number
Start date	Reference date
End date	Budget balance
Indication person	