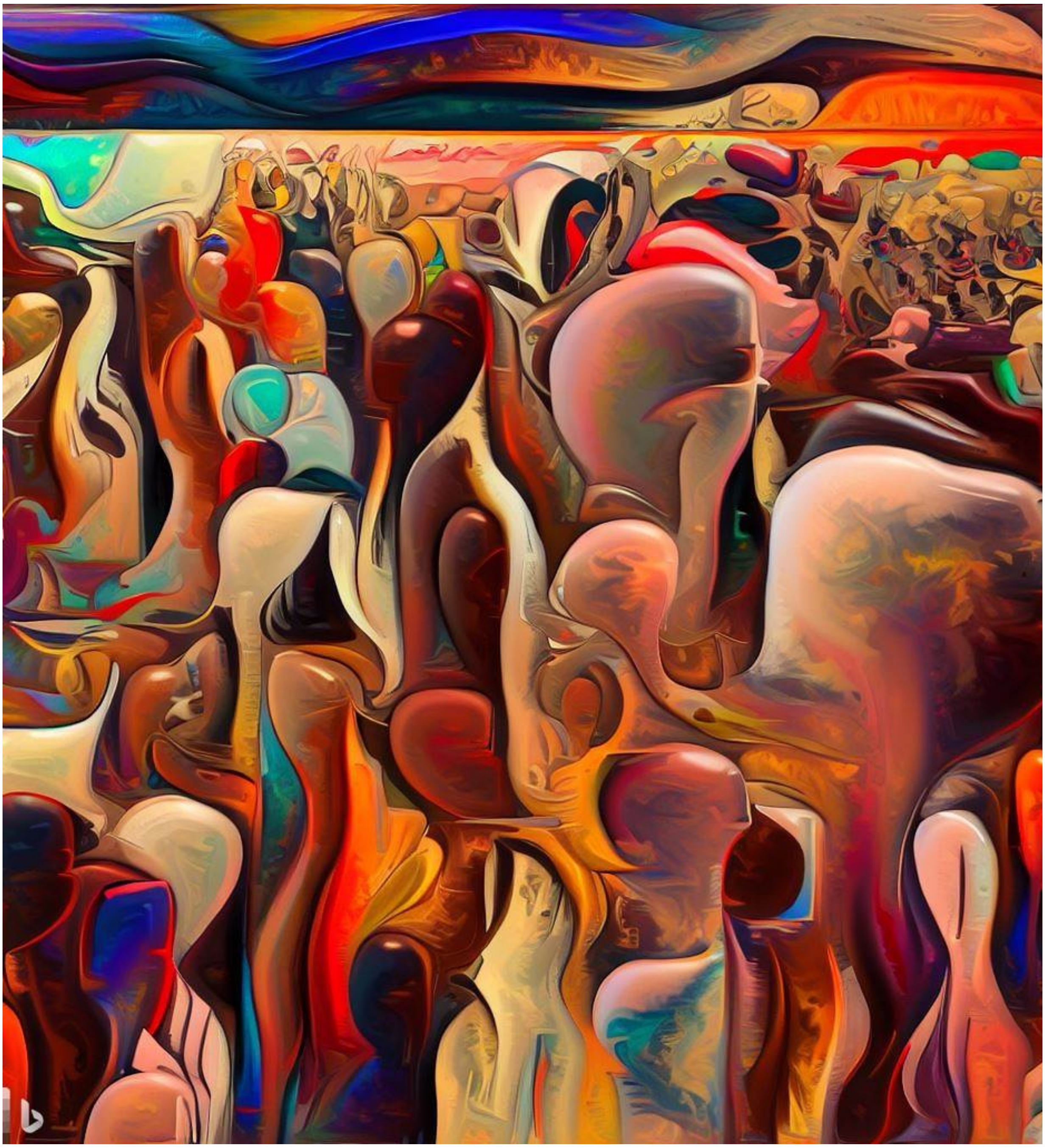


TO AI or not to AI? A responsible cooperation between civil servant and AI in decision making in the public domain.

W.E. Voskens

TO AI OR NOT TO AI? HOW DO CIVIL SERVANTS TRADE-OFF DIFFERENT LEVELS OF AUTOMATION IN LOCAL GOVERNMENTS?

A thesis submitted to the Delft University of Technology in partial fulfillment
of the requirements for the degree of

Master of Science Engineering & Policy Analysis

by

Wieger Voskens, 4582292

To be defended in public on June 19th 2023

Wieger Voskens: *To AI or not to AI? How do civil servants trade-off different levels of automation in local governments?* (2023)

The work in this thesis was made at the:

Faculty of Technology, Policy & Management
Delft University of Technology

TU Delft Supervisors:	Prof.dr.ir. N. Bharosa, Section Engineering Systems and Services
	Phd. A. Sattlegger, Section Engineering Systems and Services
	Dr. S. Buijsman, Section Ethics and Engineering
External Supervisor:	Nicolaas Heyning

EXECUTIVE SUMMARY

Introduction
The responsible implementation of AI in the public domain, especially in smaller governmental organisations such as municipalities, is a challenge that has risen to the forefront due to recent developments in the AI domain and the lack of clear regulations and policy. This study aimed to provide guidance for those willing to experiment with the implementation of AI in the public domain by identifying the main criteria that can determine the most responsible and appropriate level of automation for a decision task. The focus of the study was on hybrid human-AI cooperation, where the civil servant communicates with the AI and where the role of the civil servant needs to be defined. The research aimed to analyse the trade-offs that are made between the two main criteria when determining a responsible level of automation. The research question was: *"What trade-offs are made between efficiency and human control when determining a responsible level of automation for different decision-making processes in local governments?"*

Research Approach
The research was conducted using desk research to identify the criteria in literature and the different aspects of decision-making in the public domain where automation can be beneficial following the HACT model and different theories. Expert interviews with experts from the private, public, and scientific domains were conducted to identify and validate the findings from the desk study and to gain input for the final research method. Finally, a vignette experiment was performed, presenting four scenarios that differ in two criteria, efficiency gain and human control.

Main Findings
The findings of the study indicate that responsible implementation of AI in the public domain depends mainly on the trade-offs between the criteria *Efficiency Gain* and *Human Control*. When looking at different decision-making cases, as they were presented in the vignette experiment, context factors such as sensitivity and the impact of the decision on the individual citizen play an enormous role in this trade-off. For highly sensitive decision-making, there is a trade-off between a minimum level of human control and the amount of efficiency gain, where the latter is not considered to be the main driving force for determining a responsible level of automation. For lesser sensitive decision-making, efficiency gain is a more important criterion. Still, interestingly, there seems to be no significant difference between different levels of human control where the civil servant can or cannot impact the final decision. Thus for less sensitive decisions, a higher efficiency gain would be preferred where the level of human control does not seem as important.

While human control was deemed vital by all experts, a higher level of automation reduced human control, as AI exerts more influence over decision-making. Hybrid levels of automation featured varying degrees of human and AI control. The interviews emphasized the necessity of maintaining a certain level of human control, although this came at the expense of efficiency gain due to the lack of a complete understanding of AI decision-making processes. In cases where routing would be involved, more efficiency gain was preferred, even if it resulted in a lesser human-controlled system. The specific goals and objectives of AI implementation determine the balance between efficiency gain and human control. additional trade-offs related to efficiency gain and human control include individual impact, the complexity of cases, public acceptance, and the potential for civil servants to learn and

improve their decision-making. Flexibility and a nuanced understanding of context factors were advised when determining appropriate levels of automation.

Before implementing a responsible level of automation, it is advised to look at an experimenting or pilot phase where the level of automation can be tested, also for the civil servant to get adjusted to it. Running the AI next to the normal decision-making can also already be seen as the ″mirror″. This may, in the beginning, not result in more efficient decision-making over time, but the quality will rise, and as the civil servants become more familiar with AI and they develop more trust in using these tools, efficiency can be increased. This study suggests that there should be a fine balance where the citizen eventually benefits from the best of both worlds regarding civil servants (human) and AI in the hybrid human-AI decision-making, complementing each other. More efficient decision-making through the implementation of AI improves the handling of tickets at governmental institutions but with a level of meaningful human control where the civil servant can have more time for direct one-on-one contact with the citizen providing explanations, reasoning, and help.

Recommendations
The study concluded that AI should be seen as a supportive tool to help civil servants increase the quality of decision-making and improve efficiency with meaningful human control. Presenting AI as a mirror for civil servants enables them to learn from the AI and be more aware of their biases and mistakes, eventually leading to better social interactions with citizens. The sensitivity of the decision impacts the preferred level of human control and efficiency gain which will result in a different level of automation. In the hybrid human-AI decision-making process, a fine balance is necessary to benefit citizens with the best of both worlds, complementing each other. Finally, the role of civil servants should be to provide explanations, reasoning, and help directly to citizens, while AI can handle more mundane tasks.

ACKNOWLEDGEMENTS

This thesis marks the end of my Master's degree in Engineering and Policy Analysis at Delft University of Technology. Next to that, it is the end of one of the many phases in life where I will now spend my last days as a student in Delft. This section contains my gratitude to all those who have advised and supported me throughout the writing and research of this thesis.

First of all, I would like to thank my supervisors from the TU Delft, Nitesh Bharosa, Antonia Sattlegger and Stefan Buijsman, who have been there when I needed assistance and provided me with valuable feedback. Antonia, you have been very involved and I could always send you an email or text you whenever I had a question. Your feedback was extremely valuable and you provided me with a sounding board with whom I could spar and share ideas. Nitesh, as my chair you were obviously involved as well and I very much appreciated all the sessions that we had where I learnt from every single one of them. I would also like to thank you for inviting me to the session with the Ministry of Internal Affairs which was very interesting. As my third supervisor, Stefan, you have been extremely valuable in providing another more ethical perspective on my research. Thank you for your critical and constructive feedback and for bringing me in touch with some interesting researchers at the REAIM Summit.

Secondly, I would like to show my gratitude to my supervisor from Councyl, Nicolaas Heyning. I very much enjoyed walking around at Councyl and look back with a lot of joy on sitting down with you to discuss my thesis and also on the way that I was involved in some of the daily life business of Councyl. For this, I would also like to thank the other people working at Councyl, Caspar, Annebel, Stella and Monica. Not to forget, I would like to thank all the interviewees and respondents for their valuable input and for their time. This has been an eye-opener for me and has been very insightful for this research.

Finally, I would like to thank all the other people, family, and friends, who were always willing to help or to listen. A special thanks to Jeroen Delfos for taking so much time to sit down with me and to spar and think along with me, but also to correct me and provide me with constructive feedback. Quite some friends were somewhere in their own trajectory of conducting the research for their theses and provided me with laughter, plenty of coffee chats and new perspectives. A special thanks to Frank, Maarten, Flip, Joel, Tom, Sytze, Mees and Ricardo.

I look back with a lot of content and enjoyment and I am looking forward to all that is next!

All the best,

Wieger Voskens
June 4th, 2023

CONTENTS

1 Introduction 1
1.1 Digitalising the Public Domain 1
1.2 Research Question 3
1.2.1 Research Gap 4
1.2.2 Research Objective & Sub-Research Questions 4
1.3 Councyl 5
1.4 Report Outline 6
2 Literature Review 7
2.1 Literature on Public Domain AI Implementation 7
2.1.1 Current Knowledge 7
2.1.2 Literature Findings 8
2.2 Research Blocks 9
2.2.1 Level of Automation of AI 10
2.2.2 Criteria for AI implementation 10
2.2.3 Local Governmental Organisations 13
2.3 European Guidelines 17
2.3.1 Artificial Municipalities 17
2.3.2 Digital Civil Servants 18
2.3.3 Artificial Local Governments 19
2.4 Conclusion 19
3 Methodology 21
3.1 Research Approach 21
3.1.1 Research Scope 21
3.1.2 Strategy per Sub-Question 22
3.2 Case Study 24
3.2.1 Case Selection 25
3.3 Data Collection 25
3.3.1 Desk Research 25
3.3.2 Semi-Structured Interviews 26
3.3.3 Survey Analysis 27
3.4 Data Analysis 28
3.4.1 Stakeholder Analysis 28
3.4.2 Interview Data Analysis 29
3.4.3 Vignette Experiment 29
3.5 Conclusion 32
4 Level of Automation 34
4.1 Existing Classifications 34
4.2 Criteria for these Levels of Automation 37
4.2.1 Efficiency Gain 38
4.2.2 Human Control 39
4.2.3 Quality 40
4.2.4 Implementation Effort 40
4.2.5 Deskilling 40
4.2.6 Vulnerability 41
4.2.7 Acceptability 42
4.3 Identified Trade-Offs 42
4.4 Frameworks for Automation 43
4.4.1 Flow Chart for Level of Automation 43
4.4.2 Human-Automation Collaboration Taxonomy 44
4.4.3 Where to Use an Algorithm 46
4.5 Theories on Civil Servant Control 47

4.5.1 System Safety Theory . . . 47
4.6 Conclusion . . . 48
5 Results 50
5.1 Expert Interviews . . . 50
5.1.1 Reasons for Automation . . . 51
5.1.2 Criteria for the Level of Automation . . . 52
5.1.3 Context Factors . . . 56
5.1.4 Responsibility and Autonomy . . . 58
5.1.5 Interview Conclusion . . . 60
5.2 Vignette Experiment . . . 61
5.2.1 Survey Sample . . . 62
5.2.2 Case Description . . . 63
5.2.3 Results . . . 65
5.2.4 Survey Conclusion . . . 74
5.3 Canvas . . . 75
5.4 Conclusion . . . 75
6 Discussion and Conclusion 77
6.1 Main Findings Recommendations . . . 77
6.1.1 Hybrid Human-AI Systems . . . 78
6.2 Trade-Offs . . . 78
6.3 Meaningful Human Control . . . 81
6.4 Academic Reflection . . . 82
6.4.1 Limitations . . . 82
6.4.2 Scientific Contribution . . . 83
6.4.3 Societal Relevance . . . 84
6.5 Relevance to Study . . . 84
6.6 Future Research Suggestions . . . 84
6.7 Concluding Remarks . . . 85
A Appendix D: Value Sensitive Design 96
A.1 Value Identification through Sustainable Development Goals . . . 96
B Appendix B: Interviews 98
B.1 Interview Protocol . . . 99
B.2 Interview 1, PhD Researcher TU Delft . . . 100
B.3 Interview 2, Senior Researcher TU Delft . . . 103
B.4 Interview 3, Coordinating Policy Officer in the Domain of AI . . . 106
B.5 Interview 4, Software Develop Data Analyst . . . 109
B.6 Interview 5, Deep-Fake Expert . . . 111
B.7 Interview 6, Project-Manager . . . 113
B.8 Interview 7, Researcher in Informatics and Responsible Implementation of Technologies . . . 115
B.9 Interview 8, Policy Advisor on Ethics . . . 117
B.10 Interview 9, Project-Manager . . . 121
C Appendix C: 124
C.1 Vignette experiment . . . 124
D Appendix D: 127
D.1 Canvas . . . 127

LIST OF FIGURES

Figure 2.1 Challenges of AI, adapted from Wirtz et al. [2020] 9
Figure 2.2 The recursive VSD tripartite framework, adapted from Umbrello [2020] 13
Figure 2.3 Automation of government services, adapted from Engin and Treleaven [2019] 16
Figure 3.1 Research Strategy per Sub-Research Question 24
Figure 3.2 Summary of steps and decision points in conducting an experimental vignette methodology study, adapted from Aguinis and Bradley [2014] 32
Figure 4.1 Flow chart showing the application of the model of types and levels of automation, adapted from Parasuraman et al. [2000] 44
Figure 4.2 The three collaborative decision-making process roles: moderator, generator, and decider, adapted from Cummings and Bruni [2009] 45
Figure 4.3 A standard control loop, adapted from Laveson [2012] 48
Figure 5.1 Responsibility Triangle with their relationships 61
Figure 5.2 Vignettes on the scales of the criteria and context factor 62
Figure 5.3 Likert distributions for the four vignettes 67
Figure 5.4 Distributions per scenario for the different domains 70
Figure 5.5 Domain Averages 71
Figure A.1 VSD design process for AI technologies, adapted from Umbrello and van de Poel [2021] 96
Figure A.2 United Nations Sustainable Development Goals, adapted from UN [2022] 97
Figure C.1 Introduction to the survey (*in Dutch*) 124
Figure C.2 Example of a slider in the survey (*in Dutch*) 125
Figure D.1 Canvas 128
Figure D.2 Canvas, methods and techniques 129

LIST OF TABLES

Table 2.1 Laws and Regulations for the Use of AI by Municipalities VNG [2020a] . . . 18
Table 3.1 Interviewees . . . 27
Table 3.2 Recommendations for vignette content, adapted from Evans et al. [2015] . . . 31
Table 4.1 Levels of Automation by Sheridan and Verplank [1978] . . . 35
Table 4.2 Levels of automation by Roehl [2022] . . . 36
Table 4.3 Levels of automation specified for this case study . . . 36
Table 5.1 Reasons for Automating Decision-Making, Coded Responses 52
Table 5.2 Criteria for level of automation, Coded Responses . . . 53
Table 5.3 Criteria for Meaningful Human Control, Coded Responses . . 55
Table 5.4 Context factors that play a role in determining the level of automation, Coded Responses . . . 57
Table 5.5 Case descriptions as provided in the vignette experiment (in Dutch with translation) . . . 64
Table 5.6 scenario descriptions as provided in the vignette experiment (*translated*) . . . 65
Table 5.7 Descriptive statistics for the four scenarios . . . 66
Table 5.8 Implementation of Efficiency Gain and Human Control in the Levels of Automation . . . 73
Table 6.1 Recommendations for Main Research Blocks . . . 79
Table B.1 Interviewees Overview Appendix A . . . 98
Table C.1 Case descriptions as provided in the vignette experiment (in Dutch) . . . 125
Table C.2 scenario descriptions as provided in the vignette experiment (*translated*) . . . 126

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ML	Machine Learning
NN	Neural Networks
DST	Decision Support Tools
DSS	Decision Support System
IoT	Internet of Things
BAIT	Behavioural Artificial Intelligence Technology
DCM	Discrete Choice Modelling
ADM	Automated Decision-Making
SST	System Safety Theory
EU	European Union
EC	European Commission
UN	United Nations
CS	Civil Servant
VSD	Value Sensitive Design
PP	Parking Permit
SW	Special Welfare.
HACT	Human-Automation Collaboration Taxonomy
EPA	Engineering Policy Analysis

1 INTRODUCTION

The automation of human tasks is nothing new, even when looking at decision-making by humans that has shifted towards more automated ways, such as robots, Internet of Things devices (IoT), or Artificial Intelligence appliances such as Chat-GPT by OpenAI. Kuziemski and Misuraca [2020a] mentions the need for more in-depth analysis of the automation of decision-making especially in the public domain and regarding civil servants [Vinadio et al., 2022]. Especially in the Netherlands, where everybody is on edge when discussing the automation of human decision-making after maltreatment in the case of the 'Toeslagenaffaire'. Hood [1991] has been one of the first to discuss different levels of automation and the different possibilities of automating processes, tasks and decision-making. Including all the different stakeholders in the implementation of a certain level of automation is key for a greater understanding of accountability and responsibility aiding to a more widely supported automation of decision-making [Gesk and Leyer, 2022]. Different levels of support by automation tools such as AI require more research to be able to benefit maximally from the advantages of these new technologies. While watching over the misuse of AI resulting in HAL-like situations, the supercomputer from the science-fiction movie '2001: A Space Odyssey', where interaction between computer and human evidently fails. The level to which AI can and should overtake human tasks and decisions needs to be clearly defined to get a good working hybrid human-AI system.

1.1 DIGITALISING THE PUBLIC DOMAIN

The world is digitalising fast and governments try to provide for this need for automation and to improve the efficiency, quality and transparency of services [Lindgren et al., 2019; de Mello and Ter-Minassian, 2020]. Designing and implementing the correct strategies for these vast emerging and developing technologies by governments is a huge challenge [de Mello and Ter-Minassian, 2020; van Engers and de Vries, 2019]. Policymakers struggle with deciding on the correct moves when discussing the rise and implementation of new information technologies [Nemitz, 2018]. Especially the use of AI, IoT and blockchain brings new challenges to the table where policymakers barely have the time to decide upon new policies because the implementation of these technologies has already resulted in new challenges [Kuziemski et al., 2022; Bharosa, 2022; de Mello and Ter-Minassian, 2020; Wirtz et al., 2019]. Government Technologies (GovTech) as mentioned in the paper by Bharosa [2022], refer to these kinds of socio-technical solutions that link private organisations that operate in the technology with public sector implementation [Bharosa, 2022]. Governance and policy advice is lacking while the implementation of supportive information systems such as AI is already in place, either in pilot form or in full use [Tangi et al., 2022]. What is defined as AI for this research will be explained in the last part of this section, and is in line with the European standards and definition.

This lack of policy poses risks and issues concerning not only the privacy of users, which in the public domain implies the citizens for which the government is responsible but also the autonomy of the public services and the quality of these services, to name but a few. While the responsibility for the use of these information technologies seems to be vaguely distributed over all parties involved resulting in no responsible party in case things go wrong [Agostino et al., 2022; Arnaboldi et al., 2022]. Should it be the task of governmental institutions to safeguard this data and provide a secure digital environment for their citizens? Or is this part of the citizen's responsibility? Or maybe of the developer of the new technology? Mitrou et al. [2021] mention the different implementations of AI in most local authorities where they support decision-making by civil servants though also changing the discretionary role of public servants. AI has plenty of advantages for example improving the quality, efficiency and effectiveness of services provided in the public domain [Tangi et al., 2022; Roehl, 2022]. The use and implementation of AI, however, can still be considered a playing field where the boundaries are either not strict or have been undefined. Especially in the public domain where the question arises of how AI can help in providing better services to citizens or be more efficient regarding the role of the civil servant and their discretion.

The questions on how to responsibly implement algorithms and whether AI will not overtake and out-run us as human beings, up to a level where we do not anymore understand what we created, are key to take into account [Canhoto and Clear, 2020; Schemmer et al., 2021; Nof, 2009]. This poses an issue of explainability and equally important, it poses an issue of responsibility, where already some interesting ideas have been proposed that need to be addressed thoroughly [Elliott et al., 2021]. Governmental institutions should be getting ready for widespread implementation of decision support tools like AI, improving not only research and innovation but also administrative tasks, as is concluded by the European Commission [Tangi et al., 2022]. In order to dive into this topic, one needs to have a clear definition of AI, which differs often between scientific research. Therefore the focus of this research will mainly be on the implementation of one type of AI in administrative tasks of governmental institutions. Specifically at those governmental institutions of a smaller size where knowledge, data and other resources are often lacking and knowledge on implementation challenges is not up-to-date [Schaefer et al., 2021; Rodrigues and Franco, 2021; de Mello and Ter-Minassian, 2020; Mikalef et al., 2022, 2019].

Before continuing, the definition of AI needs to be clarified so that the reader understands what AI entails in this MSc thesis. Since this paper focuses on the implementation of AI in different levels of automation in the public domain it makes sense to follow the definition as stated by the European Commission. Following the first versions of the AI act by the European Commission, Artificial Intelligence is *"software with human-defined objectives, that generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with using different techniques"* [EC, 2019, 2021b]. Which is specified more by the EU into both general and narrow AI, where almost every AI is considered narrow, meaning these *"systems can perform one or few specific tasks that humans can do"* [EPRS, 2022]. This paper will stick to this definition, considering every software that can take over a task or multiple tasks of human beings to reach a certain goal using data that is continuously being given through its environment.

The impact of different levels of automation on the decision-making of these administrative tasks and thus on the citizens is huge and brings another enormous issue of safety [Dey and Lee, 2021]. As Dey and Lee [2021] mention, most recent incidents around the implementation of new technologies in decision-making have been caused either by insufficient training or by choosing the wrong level of automation. Understanding these levels of automation and the challenges is key to a good implementation of this technology and of a private-public collaboration [Dey and Lee, 2021].

One could state a number of criteria that the level of automation has an impact on and where it can be judged for its performance. As König and Wenzelburger [2021] mentions, there are multiple ways of designing decision-making based on algorithms such as AI, that impact these criteria differently but also result in different outcomes. The trade-offs between these criteria, which will be discussed in the next paragraph, have to be dealt with early on since these are directly linked to the operationalisation and thus the level of automation that should be implemented [König and Wenzelburger, 2021]. Wirtz et al. [2021] sum up the future research prospects regarding the use of AI in the public domain, of which a few will be addressed in this research. This research aims to provide more insights into the perceptions of stakeholders towards the implementation of different levels of automation in which human works together with AI. Also trying to provide valuable insights into the issues of accountability, responsibility, human interaction, and human values, especially in the public domain with a focus on control by the civil servant. This research aims to add knowledge also to these security issues of control by providing more insights into the impact of these different levels of automation on the criteria that actors say to be important.

1.2 RESEARCH QUESTION

The aim of this research is to explore the possibilities of AI implementation within the public domain, mainly focusing on decision-making tasks by civil servants that have a direct impact on individual citizens. And foremost to find to what extent AI can responsibly support or overtake certain decisions in a governmental institution. A clear framework needs to be set up here on the technological, judicial and ethical limitations. This research is of an exploratory nature and focuses on the trade-offs that experts make when deciding upon a responsible level of automation and on how this is perceived by civil servants. This results in the following research question:

What trade-offs are made between efficiency and human control when determining a responsible level of automation for different decision-making processes in local governments?

This research question can be divided into the following components and concepts:

- Trade-offs: Different governmental institutions (national or local) have different criteria that they value differently, implementation of AI could result in these criteria conflicting in different ways[Hood, 1991]. This is important to understand in order to implement an AI in the most accepted and responsible way possible, according to the different criteria [van de Poel, 2021]. At least to implement AI in a way that is understandable and defensible from an implementation point of view.

- Level of automation: The level of automation of AI can be seen as the level to which the AI has an impact on decision-making [Roehl, 2022]. To what extent can the AI autonomously decide upon an administrative task? To what extent is a human still involved in the decision-making? Different levels of influence that an AI has will be discussed and defined, ranging from full automation to monitoring and introspection.

- Local governments: This part focuses on smaller governmental institutions, such as small municipalities, where the implementation of an AI faces other problems than larger governmental bodies. For this, it is also important to note that the organisational level that will be examined is focused on what level of AI should be implemented instead of how this will be implemented. Thus focusing on the policy advice and not on the implementation itself.

Sub-questions of this research will focus on these main concepts as mentioned above. The willingness of people and organisations to responsibly use AI and implement a level of automation, while taking into account the impact on citizens is a main challenge here. This challenge is addressed by looking at the context factors and the trade-offs between the conflicting criteria, that arise when deciding upon a certain level of automation.

1.2.1 Research Gap

Up to now, most research into the implementation of new information technologies in the public domain has focused on what type of digitalisation tool should be implemented and on the issues that come forward with implementing such new technologies. For example choosing between a decision-support tool, neural networks (NN), or machine learning (ML) [Delfos et al., 2022]. In the private domain, however, plenty of new technologies have already been further developed and implemented in ways that are not yet seen in the public domain [Reis et al., 2019]. Where quite some research has been written on the collaboration between the public and private domain in implementing new technologies, more is needed [Khan et al., 2020; Bharosa, 2022]. Reis et al. [2019] stresses here the importance of understanding decision-making tasks by civil servants in the public domain and the importance of understanding the implications of governance on new information technologies in the public domain. Wirtz et al. [2021] among others have mentioned the importance of more in-depth empirical research into the public domain, especially focusing on real-life cases and AI applications. The aim of this research is to provide a holistic understanding of the implementation of AI in the public domain in the Netherlands, specifically on the level of automation of administrative tasks performed by smaller public authorities such as municipalities.

The sub-questions support the main research question in that they divide the topic into smaller blocks that could be answered separately but are inevitably intertwined as can be concluded from the later two sub-questions. Here the sub-questions come together, aiming to answer the main research question and to provide valuable insights for governmental institutions on the criteria that impact the level of automation for certain administrative tasks and how these criteria can be addressed and how they impact each other.

1.2.2 Research Objective & Sub-Research Questions

As stated his research aims to provide insights into the criteria and trade-offs that determine a responsible level of automation when automating decision-making by civil servants through artificial intelligence within smaller public organisations, such as municipalities, in the Netherlands. This research is of a qualitative nature and aims to describe and understand the reasons behind these trade-offs and to explore the different levels of automation. To achieve this, the research question is supported by three sub-research questions as presented below.

Research question: *What trade-offs are made between efficiency and human control when determining a responsible level of automation for different decision-making processes in local governments?*

Sub-research questions

Supporting the main research question, the following three sub-questions have been defined to give a well-layered and properly founded answer to the main research question. These sub-questions all address different parts of the knowledge gap as further explained in chapter 3. The order of the sub-questions is needed to answer the main research question though it also aligns with the methodology as will be explained in section 3.1.2.

1. *What criteria do experts see in automating direct decisions making for citizens, regarding different context factors?*

   Before being able to analyse the trade-off that experts make it is important to first get insights into the criteria that they think are important in determining a certain level of automation. Besides the criteria, it is important to get an idea of all the restrictions around the implementation of AI in the public domain, considering the judicial and ethical frameworks that are already present. For this research, we focus on contextual factors that impact the trade-offs between the criteria.

2. *How do experts make the trade-off between human control and efficiency gain in deciding on a level of automation by looking at different context factors resulting from the interviews??*

   Following the criteria from the previous question, the trade-offs between efficiency gain and human control can be investigated, when looking at different cases. Besides the criteria, it is important to clearly state the scenario and define the contextual factors in order to have all the civil servants interpret the case the same. We focus on the trade-offs that experts see but regarding the civil servant perspective as they play a role in the human control criteria.

3. *How do the trade-offs help in determining a responsible level of automation for different decisions, regarding sensitivity and impact on citizens?*

   After identifying the trade-offs, the reasons behind these trade-offs and the train of thought of the experts regarding the role of the civil servant need to be identified for the policymakers to be able to use these insights in implementing a responsible level of automation. Combining the information from these three sub-questions provides insights for the main research question that could be beneficial for a safer, more responsible and more accepted way of implementing AI in decision-making by civil servants in municipalities in the Netherlands.

1.3 COUNCYL

Councyl is a spin-off of the Technical University of Delft. They have developed an AI algorithm that can support the decision-making process in for example the public domain. It can be implemented in multiple other domains but for this research, Councyl has provided this topic, where understanding needs to be gained from the collaboration between private and public organisations. In digitising the public domain these collaborations bring enormous advantages but also risks and management challenges that must be addressed [Bharosa, 2022; Reis et al., 2019]. Councyl is interested in how they can assist governmental institutions by showing them what criteria are more important in determining the level of automation, and what trade-offs are made. This research will contribute to this by giving insights into the trade-offs that experts make from a civil servant perspective when given different levels of automation, focusing on the main criteria as indicated by these experts. These insights will be useful for both the public domain and private organisations and add to scientific knowledge on criteria and trade-offs for levels of automation while also creating advice for policymakers on the responsible implementation of these different levels of automation.

1.4 REPORT OUTLINE

The remainder of this report will have the following outline. In Chapter 2, a brief literature review is conducted building on the introduction and the foundation for this research is provided. The research approach and the methodology will be presented in Chapter 3. Chapter 4 will show the results from the desk study and will dive into the different aspects of the concept of level of automation forming the theoretical operationalisation for this research. Subsequently, Chapter 5 will present the results and the findings from this research. Chapter 6 will discuss the main findings, limitations, recommendations and future research possibilities.

2 LITERATURE REVIEW

This chapter will elaborate on the literature review that has been conducted on this topic of responsible AI implementation in the public domain. After elaborating on the reviewed literature, the different research blocks that are present in the main research question will be discussed before expanding on a more in-depth analysis and explanation of the demarcation made for this research. Finally, a brief notion will be made on judicial conditions and current laws and regulations on the use of AI, as it is important to be aware of this constantly developing topic, though this will not be the main concern of this research. After the clarification of these theoretical concepts, chapter 3 will elaborate on the methods used to answer the research questions in this research.

2.1 LITERATURE ON PUBLIC DOMAIN AI IMPLEMENTATION

Plenty of examples can be found in the literature on AI implementation in the public domain where the algorithm is allowed to take a decision or at least to advise on the decision [Goodrich and Boer, 2000; Wang et al., 2019; Martinho et al., 2021]. Famous examples where AI was not well implemented are the use of AI in facial recognition used by law enforcement, and so-called suspect profiling, which turned out to be extremely discriminatory and privacy intrusive in some cases [Raaijmakers, 2019; Ojamo, 2021]. Or examples in the health care systems that bring issues of privacy and fairness that have been feeding the discussion on ethical acceptance of AI implementation [Karimian et al., 2022; Sandeep Reddy, 2019; Martinho et al., 2021]. However, a lot of examples also show how it can work or at least show the potential that AI has in supporting and assisting humans in complex or recurrent decision-making [Qin et al., 2018; Cockburn et al., 2019]. Different levels of influence of AI on human decisions have been allowed in these examples, but no guidelines have been given on how AI should be implemented properly and foremost responsibly. As Gotthardt et al. [2020] mention, organisations are lacking behind while the capabilities of these technologies are expanding and rapidly growing. Policymakers need to be aware of the full risks but also of the possibilities that come with AI, thus the need for guidance is growing [Gotthardt et al., 2020; Karimian et al., 2022]. This need has maybe even become more prominent due to the newest hype with ChatGPT created by OpenAI [Sundar, 2023; Oviedo-Trespalacios et al., 2023].

2.1.1 Current Knowledge

Apart from companies and institutions, laws and regulations are also lacking behind, failing to provide proper guidelines for AI implementation [Nemitz, 2018]. Especially in the public domain, where the use of AI is relatively new but vital to the ever-growing demand for good, efficient, transparent and explicable service, this research can help to provide guidelines. Governments have been implementing decision support systems and AI, though one-third of those are still in a pilot phase [Tangi et al., 2022]. As Tangi et al. [2022] state, if the implementation of AI systems further develops, a focus on AI risks and mitigation measures is needed in order to maintain the fairness of these systems in fighting inequality and discrimination [Tangi et al., 2022]. Governments can use AI as a support tool in plenty of decisions they have to make and tasks that they have to fulfil. The question lies in the reliability of the software, on privacy-preserving, on the responsibility and the level of automation of AI on human decisions [Roehl, 2022; Schaefer et al., 2021]. Since AI can also do more than just support decisions it is important to understand when what level of automation on human decisions is wanted or needed. This can differ from introspection to full automation of decision-making by the AI according to Roehl [2022]. From a scientific point of view, this research can benefit to the understanding of the risks of these different forms of using AI and how we could address these risks and deal with them accordingly. An assessment framework, that helps and advises in determining to what level algorithmic decision-making is acceptable and responsible, is missing. Not every AI is the same and therefore not every implementation of AI should be the same. The way AI should be used depends on different criteria but also on contextual factors and on the goal of the automation process. A framework to determine the level of automation of AI on human decisions could be very beneficial for future AI implementation considering all conditions and complexities that play a role. Aiding both public organisations as well as the AI developers and companies. Hence, the goal of this research is to provide guidance for policymakers looking to experiment with algorithmic decision-making, focusing on one specific sort of AI. Since different AI implementations and usages require different policy measures and may use different levels of automation and are suitable for different types of tasks.

2.1.2 Literature Findings

This section shows the most important findings from the literature research conducted for this proposal in order to identify the knowledge gap and to find all possible research directions. A few conclusions can be made from this literature research regarding the implementation of supportive decision tools such as AI in the public domain.

- AI has been implemented by multiple public organisations and by governments. The risks and threats have not been thoroughly identified by these institutions and thus there is a lack of general understanding and way of dealing with these issues in the public domain. Especially when considering all the different types of AI algorithms and different purposes. Although plenty of reports and acts are now being published on a European level, for smaller governmental institutions it is still grasping at straws. This research aims to provide this need for handles at a more local level of responsible AI implementation.

- The frame that focuses itself on the use of AI is not fully defined resulting in uncertainties and different perceptions of what is allowed and what is best for the users (civil servants) or the citizens. It is therefore important to look at a specific AI implementation at a specific form and level of government to be able to set up guidelines for proper implementation.

- A third point that may be the main driving force behind this research, is that little is written on the level of automation of AI in the decision-making process of civil servants in the public domain. Different laws, regulations, conditions, ethical reasons etc. will play a role in every different decision that has to be made by an individual working in the public domain. For these decisions that can be automated to a certain level, it is important to find out what criteria impact this level of automation and how this level of automation can benefit the goal of providing quality, efficiency and transparency in public service.

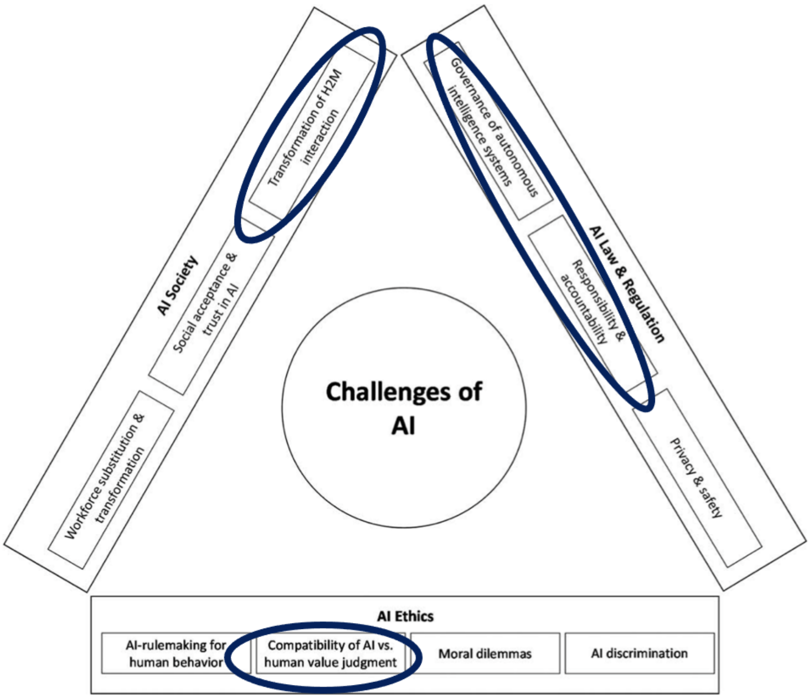

Challenges of AI
AI Society
Transformation of H2M interaction
Social acceptance & trust in AI
Workforce substitution & transformation
AI Law & Regulation
Governance of autonomous intelligence systems
Responsibility & accountability
Privacy & safety
AI Ethics
AI-rulemaking for human behavior
Compatibility of AI vs. human value judgment
Moral dilemmas
AI discrimination

Figure 2.1: Challenges of AI, adapted from Wirtz et al. [2020]

Figure 2.1 shows the challenges of AI as stated by Wirtz et al. [2020]. The circles show the challenges that will be addressed in this research based on the literature study and resulting from the knowledge gap. In the left corner, the social acceptance and trust of AI is an important part of the correct implementation of AI in society. But this research focuses mainly on the human to machine interaction and how the different levels of automation give different ways of interacting between humans and machines. In the context of law and regulation, this research tries to provide concrete advice on policymakers experimenting with AI in the public domain thus supporting governance decisions. With the different levels of interaction between humans and machines important aspects of responsibility and accountability will also be elaborated on. In terms of AI ethics, many different aspects will have to be taken into account when diving into the implementation of AI in the public domain. This research focuses on the level of automation and the interaction of the human and the machine thus focusing on the compatibility of those two parties regarding value judgement.

2.2 RESEARCH BLOCKS

This section will elaborate on and show the theoretical frameworks for the different research blocks as they have been introduced in the introduction. First, the level of automation will be discussed and the literature that has been written on this topic. Next, the criteria that determine the strategy for a certain level of automation and the main trade-offs that have been found in the literature will be discussed. The level of automation will be extensively studied in chapter 4, where the criteria and trade-offs will be operationalised for the rest of the research. Finally, the literature on the implementation of AI in the public domain and its difference from the private domain will be elaborated on, focusing on the public role and the role of civil servants. A notion of safety from a meaningful control perspective will be placed here since this could be relevant to the role of the civil servant.

Before diving further into the topics of this research, it is important to briefly mention the herd of elephants in the room when talking about algorithms and especially about AI. Plenty of research has been done on the implementation of AI, its challenges and its potential. The number of papers on responsible AI will probably have increased enormously since the start of this research, especially with the help of ChatGPT that has made its arrival a few months ago. Some of these debates on AI will not be considered in this research although these are obviously very important to keep in the back of our minds when thinking about experimenting with AI, especially in the public domain when the outcome can have direct implications for individual citizens. This is merely a notion and a disclaimer that not every aspect and problem of AI implementation will be focused on. Bias in data, for example, is a valid argument that needs to be considered when thinking about AI implementation but this is not seen as the main point of interest in this research. Here, the focus lies on the implementation of a responsible level of automation and how certain criteria and trade-offs between these criteria can determine this responsible level of automation in different decision-making scenarios. It is expected in this research that with, the upcoming AI ACT from the European Commission, the AI systems that are accepted and are on an acceptable and tolerable level in the pyramid of risk need to be given a responsible level of automation depending on the task it is used for [EC, 2022; Yakimova and Ojamo, 2023; EC, 2021a].

2.2.1 Level of Automation of AI

The different levels of automation mentioned in existing literature differ per scientific paper, per purpose, or per branch. Not to mention how different levels of automation can have very different impacts and ripple effects for different cases [Mehmood et al., 2023]. With ripple effects, the unintended consequences that occur when implementing an automation tool in a certain way for a certain task are meant. Different levels of automation vary on the scale from no AI influence to full automation where AI has full control of the decision and no human being is included in the decision-making process [Gulenko et al., 2020; Endsley^2 and Kaber, 1999]. Shneiderman [2020] elaborates on the 10 different levels of automation defined in 1978 by Sheridan and Verplank [1978], where more automation comes at the cost of human involvement. This was one of the first scientific research done on different levels of automation of human tasks with the rise of new technologies. Shneiderman [2020] stresses the importance of having a human-centred design on these different levels of automation since full automation by technology does not mean the AI will be trustworthy, reliable or even safe. Which was stressed at the first summit on Responsible AI in the Military Domain (REAIM) by Callamard [2023]. Different visions on the rise of AI exist, and whether AI should be seen as another agent in the decision-making system of governmental bodies [Tangi et al., 2022; Maragno et al., 2022]. Where Shneiderman [2020] and Johnson et al. [2014] underline the importance of human-centred bn AI to ensure safety where algorithms support human decisions. Gesk and Leyer [2022] mention the need for future research in the acceptance of stakeholders of different levels of automation and hybrid forms of human-AI interaction and decision-making in the public domain. Gaining insights into the perceptions of all the involved stakeholders on the different levels of automation of AI in the decision-making process is key. Chapter 4 will focus more on the level of automation and the criteria and trade-offs that will be first introduced in the next two paragraphs.

2.2.2 Criteria for AI implementation

Certain criteria can help in determining the decision for a certain level of automation, but how the importance of these criteria differs per decision-making task and per level of automation is not always clear. This is due to the fact that the implementation of new technology can bring about changes in values and criteria which are not often included in these collaborations of public organisations and engineers [Yoshioka et al., 2021]. As van de Poel [2022] mentions, changing values due to the emergence and implementation of new technologies is an existing challenge that needs to be well addressed in order to responsibly implement these new technologies [van de Poel, 2022]. Trying to predict those changes is a way of trying to deal with these existing values and the trade-off between those values, but also the arrival of new values when implementing a new technology [van de Poel, 2022; Niet et al., 2022]. Values can be defined as that what an individual or a group of individuals finds important in life [van de Poel, 2022]. The values for AI implementation that have been discussed in the literature will come forward in the next paragraph. Values have been used as a way of measuring and analysing certain organisational structures and strategies in order to best implement new technologies or strategies [Hedström et al., 2011; Mendoza et al., 2022]. Several tensions are discussed between the values of the emergence of new technology. This is a relevant analysis, especially in determining what type of algorithm you want to implement. Value trade-offs can help in determining whether a complex AI, simple AI, decision support tool, or Neural Network (NN) is in place for supporting decision-making Delfos [2022].

Existing literature discusses values that are important at the implementation level of an AI algorithm. Several values that have been discussed extensively are quality, efficiency, transparency, comprehensibility, explicability, fairness, etc. [Bannister and Connolly, 2020; Yigitcanlar et al., 2022; Rodrigues and Franco, 2021; de Mello and Ter-Minassian, 2020; Tangi et al., 2022]. Values have been previously defined as ideals and are suggested to have an aspect of moral good, with often different and competing philosophical conceptions [Muhlenbach, 2020; van de Poel, 2021, 2022; Wilson, 2022]. Taking these (public) values into account can be very beneficial for determining a safe, responsible and meaningful implementation of AI, especially in the public domain, where they strive to give equal care to every individual citizen. The fairness of those systems is however dependent on measures that make sure that AI implementations are not discriminating or unequal [Tangi et al., 2022]. It is however very difficult to clearly define or measure these values as they are often subjective and can easily change over time [van de Poel, 2021, 2022]. This research, therefore, aims to provide a more easy to quantify and measure way of looking at certain criteria that can represent these values (partially).

This will also need to be done in this research to understand the different types of criteria and to understand the trade-off between some of these criteria when determining a level of automation. Not to forget the role of the new technology, in this case, the implementation of AI in public decision-making by civil servants. A distinction, however, is to be made between some concepts. In the literature, there is an understanding of values, requirements, criteria and impact. All four of these concepts are important to take into account when looking at the implementation of a new technology which will have a big influence on many actors and stakeholders involved. However, in literature and also in practice these concepts sometimes overlap partially or fully. A distinction could be made with values and requirements on the one hand, of which most have to be defined before implementing a new technology, and criteria and impact on the other hand, which determine the level of and the way of implementing a new technology. Plainly put, the first two could be seen to answer the question of what form of technology is wanted, regarding what is considered important in a particular scenario or domain, and the later two answer how to implement this technology in order to achieve the values by looking at the criteria for this implementation. This research focuses on the latter though it is important to understand and be aware of all concepts. The following sections and chapters will focus on the criteria that play a role in determining not *what* tool to use to automate decision-making, but *how* to use a tool to automate decision-making by determining the level of automation and the way that humans and algorithms can responsibly cooperate. After a certain tool has been determined it is important to look at the level of automation that is for a particular task. Now that the different goals of trading off these values for the type of algorithm and trading off the criteria for the level of automation have been touched upon, we can dive a little bit deeper into the trade-offs that have been made in previous literature regarding values criteria.

Trade-offs in Automating Decision-Making

This research will focus on the trade-offs that are made when determining a responsible level of automation for different decision-making tasks. Public decision-making is a task conducted by civil servants and case workers who deal with all kinds of values, norms and criteria such as quality, accountability, reliability, the obligation of reason-giving, equality of treatment and the principle of proportionality [Roehl, 2022]. Local governments face the challenges of determining what level of automation is best and how different levels of automation result in conflicting criteria. Values can differ from the governmental institution's perspective but also from the AI's perspective and the citizens' perspective. Values however do not have a clear-cut definition due to their philosophical character [**?**]. Looking at the implementation of AI and at the use of different levels of automation of AI in the public domain, it is not necessarily a case of what is morally good (values), but more of what is the wanted impact of this strategy for AI in a certain context. This required impact can however be a specification of certain values at a higher level, but more easily measurable. From now on, this research will focus on the required impact in a certain context wanted by the stakeholders regarding different levels of AI automation, the wanted outcome can be measured over certain criteria. The main driving force of automation, especially in the private domain, is an increase in efficiency [Hood, 1991]. While this is definitely an advantage of automation through AI, especially for repetitive tasks, the level of automation and human autonomy could on the other side be an issue. The more a task is automated through an AI, the less the human has an impact or influence on the decisions made by the AI, following the scales by Hood [1991]; Cummings and Bruni [2009]; Sheridan and Verplank [1978]. A complete list of criteria will be established in chapter 4 to be able to make complete trade-offs when determining a responsible AI implementation.

Value Sensitive Design

The Value Sensitive Design (VSD) is a design method that could be used in determining values that are important to take into account when designing the implementation of a new technology [Umbrello, 2020]. The three domains of investigation that this VSD consists of include a *conceptual investigation*, a *technical investigation*, and an *empirical investigation*, as can be seen in figure 2.2. Similarly one could think of identifying criteria that are important for determining a level of automation.

With the implementation of AI, especially for the usage of civil servants in providing service to citizens, these three domains need to collaborate to make sure that responsible human-AI cooperation can be established. The identification process of criteria in this research has been inspired on this VSD model following a conceptual and empirical investigation. The trilogy has also been of interest since this research includes the visions of three domains. The first domain, the public domain, has knowledge of the tasks that this AI will overtake and of the civil servants, about the values and the value trade-offs that should be made for different decision-making processes (conceptual investigation). The private domain often delivers a technical tool that may be implemented by the public domain, here lies knowledge of the

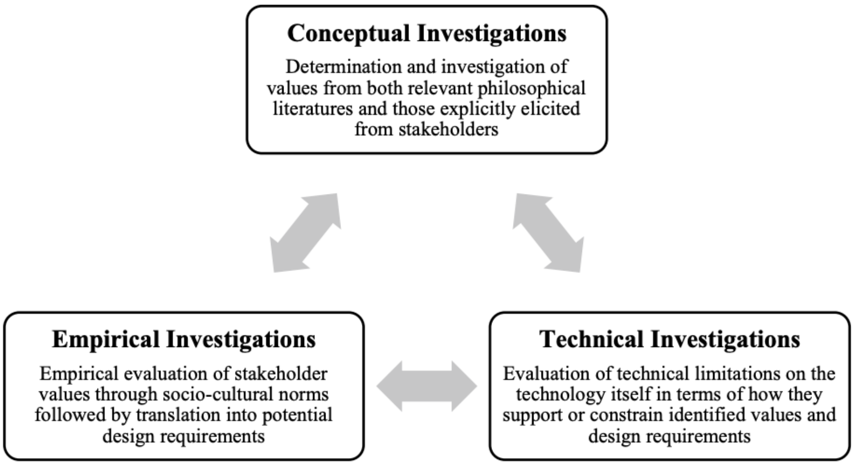

Conceptual Investigations
Determination and investigation of values from both relevant philosophical literatures and those explicitly elicited from stakeholders
Empirical Investigations
Empirical evaluation of stakeholder values through socio-cultural norms followed by translation into potential design requirements
Technical Investigations
Evaluation of technical limitations on the technology itself in terms of how they support or constrain identified values and design requirements

Figure 2.2: The recursive VSD tripartite framework, adapted from Umbrello [2020]

technical boundaries regarding the values and design requirements (technical investigation). Though this is not part of this research as the goal is not to design a system but merely to provide guidance and insights. And lastly, the scientific domain provides knowledge and empirical evidence on these values and the sociocultural norms (empirical investigation). These three domains overlap in some aspects of implementation and knowledge and should not be seen in a hierarchical order but more as intertwining with and depending on each other as well as the investigation domains from the VSD [Umbrello, 2020; Umbrello and van de Poel, 2021]. Wiesmüller et al. [2023] has elaborated on the responsible implementation of AI in businesses also regarding the three main domains present in this situation, where the only difference is the implementation domain that changes from private to public in this research. This, however, is a huge difference regarding the implementation of AI due to legal requirements for transparency and explainability and social requirements of equality and fairness for public organisations [Lindgren et al., 2019; de Mello and Ter-Minassian, 2020].

2.2.3 Local Governmental Organisations

When looking at the implementation of decision support tools such as AI in the public and private domain, huge differences can be seen of which arguably the most important are the clients of the public domain, the citizens and the way governmental institutions have to care for its citizens [Dey and Lee, 2021; Reis et al., 2019]. Whereas private companies and organisations focus on providing what their customers need and thus creating as much revenue as possible through developing and implementing new technologies [Mikalef et al., 2022]. Efficiency gain and reducing costs are the main drivers for automation in the private domain [Parasuraman et al., 2000]. Implementing AI here has a higher focus on increase in efficiency since this results in more profit whereas for public institutions the care of the citizens and the quality and safety could be considered as more important where factors such as trust and transparency play a huge role [Muhlenbach, 2020]. Within the public domain, different entities deal with issues related to the digitalisation of society and therefore need to adjust fast. On a European level, the first guidelines are being written, and implementable mainly on a national level by governments [de Mello and Ter-Minassian, 2020; EC, 2021b]. But when we zoom in on the smaller governmental bodies such as provinces, municipalities, or executing governmental organisations, it is harder to define the exact rules that are applicable to them [Methnani et al., 2021]. Digitalisation may happen (relatively) fast in large municipalities and cities that have a whole department focusing on digitalisation, but for smaller, more remote municipalities or smaller public organisations, these innovations may be hard to follow [Mikalef et al., 2022]. Resulting in unequal service and care for citizens living in smaller and more remote municipalities compared to those living in large, influential municipalities and cities [Combes et al., 2012].

Governments are trying to catch up and provide as much openness and transparency to the citizens as the new law on open government (*'Wet open overheid'*) in the Netherlands and a new bill that is coming up on a digital government (*'Wet digitale overheid*) to build on the citizen's trust in a faster digitalising government [Rijksoverheid, 2022b, 2023c]. Not to mention the impact of the the fast deployment of Chat GPTA on these developments, which resulted in plenty of administrative issues and challenges for example schools, educational institutions and universities, due to the lack of policy [Roose, 2023]. As Kuziemski and Misuraca [2020a] mention, it is key to "govern algorithms while governing *by* algorithms". Resulting from the previous sections the following observations are made based on the role of public institutions and the values trade-offs of implementing AI, following from the literature that has been examined. These observations will be tested in this research, trying to get a more complete picture of the criteria and the trade-offs.

O1: Human autonomy and quality of service are the main criteria that determine the level of automation in automating decision-making by civil servants in the public domain in the Netherlands.

O2: Cooperation between the scientific, public and private domains is needed to identify the criteria and responsibly implement AI in hybrid human-AI systems with meaningful human control.

The Netherlands

The Netherlands is an interesting case to look at as it is the third most digitalised country within the European Union (EU) [Dieter O., 2022]. It is one of the most developed countries within the EU and is a front-runner when it comes to digitalisation and new technologies, being highly innovative and with plenty of start-ups popping up every day [NordicHQ, 2023]. Also in the domain of AI or automation tools. Looking at the criteria for AI implementation the values and norms that are present in the Netherlands could be slightly different from other countries within the EU, though overall they could be considered fairly equal for the EU. This results in more easy to generalise results to the rest of the EU. Since the Netherlands is so highly innovative, plenty of start-ups that work with AI are trying to find ways of implementing this in the public domain. A lot is happening when looking at different pilots that are running and how the government is trying to organise this Algoritmeregister [2020]. This makes the Netherlands a relevant, interesting and relatable case.

Civil Servants

A lot has been written on the *technological* aspect of the implementation of new technologies in the public domain and also on the need for addressing the societal challenges that come with values such as privacy, transparency, responsibility, fairness, etc. [Engin and Treleaven, 2019; König and Wenzelburger, 2021; Wirtz et al., 2020]. As became clear from the previous sections research has shown how difficult it is to quantify these values and make them more concrete. Needless to say these values and the perspectives on these values and their trade-offs can differ from those of the policy maker, the company providing the AI or the citizen. The perspective of different stakeholders differs. This can be considered the same when considering the criteria for determining the responsible hybrid form of human-algorithm cooperation. It is therefore important to dive into those stakeholders, focusing on policymakers, civil servants and citizens. Citizens should be more included in the decision-making processes of governmental institutions, but they won't be the main focus of this research [Rovers, 2022]. The acceptance and willingness to adopt a new technology or a new form of GovTech or e-governments has been a topic of interest for many years and plenty of research has been done on this aspect [Carter and Bélanger, 2005; Gesk and Leyer, 2022; König, 2022]. Gesk and Leyer [2022] stress the importance of future research into the perceptions of other stakeholders such as the operators, the civil servants, and into the different levels of automation that arise in hybrid human-computer decision-making. This research aims to focus on the trade-offs and decisions that civil servants would make for deciding a certain level of automation of their tasks. Current policy focuses more on how we should move towards decision-making *"by"* AI, and less on working *"with"* AI where the civil servant in the public domain plays an important role [Kuziemski and Misuraca, 2020a]. How can we for example increase efficiency and accuracy, while being transparent and not letting a machine have full autonomy on morally complex decisions which can impact the trust of citizens in the government? Especially when wrong decisions are being made in the lives of individuals and no explanation can be provided [Wirtz et al., 2019]. It is therefore highly relevant to dive into the perceptions and the role of civil servants when considering hybrid forms of decision-making in the public domain between the algorithm and the civil servant. This comes back to the concept of 'meaningful' human control, where transparency, responsibility and accountability need to be correctly addressed.

The role of civil servants is crucial since they have a certain level of responsibility and accountability in the current decision-making process, and this might also be needed when working together with an algorithm. Therefore civil servants are considered the main stakeholder in this research, though all stakeholders do need to be involved in the decision-making process on these levels of automation and in the design phase of these (semi-)automated governmental systems. Figure 2.3 shows where automation can support governmental tasks and decision-making by Engin and Treleaven [2019]. This research focuses mainly on the implementation of AI in different levels of automation in the first two sectors, where AI aims to support civil servants in deciding on public services. Where Engin and Treleaven [2019] mainly focus on the technological aspect, this research strives to focus on the combination of technology and expertise of different domains while taking into account the societal impact, providing guidance and advice to policymakers. The role of the civil servant and the impact that the implementation of AI may have on this role is therefore crucial. For example, discretion by civil servants may be seen as both a blessing and a curse, but the implementation of AI will most likely change the way this operates now.

A lot has been written on the discretion of civil servants in their (administrative) decision-making and about their drive and will to deliver as good as possible service for the citizens [Bullock, 2019]. Discretion of civil servants is considered as the freedom and latitude of interpretation and of decision-making on the implementation of the governmental policy when situations become complex and ask for some human intervention and interpretation [Bullock, 2019; Mitrou et al., 2021]. Automating decision-making in the public domain can result in having less discretion which can be both beneficial and negative, as will be explained here. The existence of concepts such as *automation bias*, *confirmation bias* and the fact that humans are not irrational, make automating decision-making can be very interesting and helpful. Though, on the other side, it can also emanate in an AI implementation that does not solve these bias issues [Bullock, 2019; Selten et al., 2023]. Automation bias entails the overestimation of an algorithm such as AI, where the civil servant would be trusting too much in the rationality of an AI algorithm [Selten et al., 2023]. Confirmation bias has been extensively researched by Selten et al. [2023] and implies

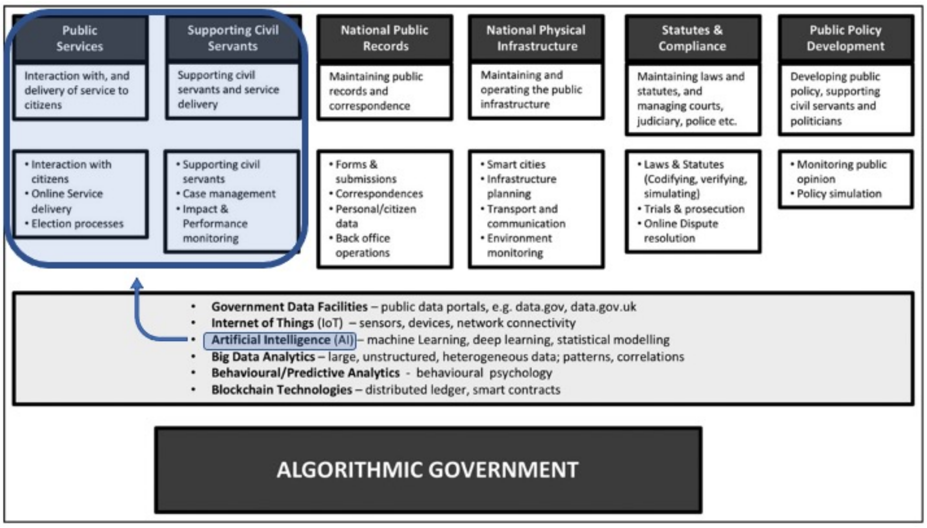

Public Services
Interaction with, and delivery of service to citizens
• Interaction with citizens • Online Service delivery • Election processes
Supporting Civil Servants
Supporting civil servants and service delivery
• Supporting civil servants • Case management • Impact & Performance monitoring
National Public Records
Maintaining public records and correspondence
• Forms & submissions • Correspondences • Personal/citizen data • Back office operations
National Physical Infrastructure
Maintaining and operating the public infrastructure
• Smart cities • Infrastructure planning • Transport and communication • Environment monitoring
Statutes & Compliance
Maintaining laws and statutes, and managing courts, judiciary, police etc.
• Laws & Statutes (Codifying, verifying, simulating) • Trials & prosecution • Online Dispute resolution
Public Policy Development
Developing public policy, supporting civil servants and politicians
• Monitoring public opinion • Policy simulation
• Government Data Facilities – public data portals, e.g. data.gov, data.gov.uk
• Internet of Things (IoT) – sensors, devices, network connectivity
• Artificial Intelligence (AI) – machine Learning, deep learning, statistical modelling
• Big Data Analytics – large, unstructured, heterogeneous data; patterns, correlations
• Behavioural/Predictive Analytics - behavioural psychology
• Blockchain Technologies – distributed ledger, smart contracts
ALGORITHMIC GOVERNMENT

Figure 2.3: Automation of government services, adapted from Engin and Treleaven [2019]

that civil servants will be more likely to adopt the decision by an algorithm when it supports and confirms their decision. Discretion is therefore a very important aspect to take into account, though it will not be the main focus of this research since quite some literature has elaborated on this topic, where new technologies have been shown to improve human discretion [Bullock, 2019]. However, with the rise of new technologies like AI in more cognitive tasks, such as the decision-making over sensitive cases by civil servants deciding on the lives of individual citizens, these discretion issues have not been fully solved yet and this level of human control and human intervention can be very pleasant to have. This research aims to provide in making insightful how discretion can also be improved through different levels of automation and hybrid forms of human-algorithm cooperation for different types of scenarios where at times less or more discretion may be needed. As de Boer and Raaphorst [2021] also state, less complex tasks can be more easily automated since the level of discretion is of less importance, and not much distinction exists in the way civil servants would decide on these tasks. For this research, the aim is to give advice on how the wanted or preferred level of human control by the civil servant, and thus discretion, can differ per situation and how this should play a role when determining a responsible level of automation.

Apart from the above-mentioned reasons, good implementation of AI is highly dependent on the way that it is used and how the operator uses and understands it [Engin and Treleaven, 2019]. There is a gap between the computation and creation of the AI and the eventual public purpose that it could be used for. Thus a greater understanding should be created through training and educating existing workforces on how to cope with new technologies. This research aims to set a first step in understanding the new role of civil servants and their perceptions in the use of new technologies, as they have the field experience and knowledge of these decisions where their expertise is again needed for AI. A trade-off in human-AI interaction here is inevitable. This research uses the vision and expertise of domain experts in applying and implementing AI in the public domain to identify this new role for civil servants. As has been discussed in earlier research, street-level bureaucrats tend to trust an AI algorithm more if it confirms their own beliefs rather than when it is contradictory to their judgement [Selten et al., 2023]. Street-level bureaucrats are the civil servants in public organisations who are in direct contact with the citizens. Based on this mention of a new role of civil servants and thus a different form of human control and based on the remarks from earlier discussed topics, the following two observations from the literature can be made.

O3: AI can be used to improve the civil servant's discretion for different decision-making processes, which can contribute to a higher quality of decision-making.

O4: Researchers acknowledge a new role of civil servants where human autonomy is more important for responsible AI implementation. *Where for example, more sensitive decisions require more discretion and more (meaningful) human control by civil servants.*

2.3 EUROPEAN GUIDELINES

Plenty has been written on a European level in the past 5 years on the "technological storm" that is coming our way if we are not already in the midst of it [EPRS, 2022]. One of the biggest new regulations from the European Commission may have been the General Data Protection Regulation (GDPR), which companies have to abide by since 2018, and deals mostly with the issue of data and how to deal with the enormous amount of data that is being generated of every individual [EC, 2020]. As stated on the website of the European Union, they strive to become "*a world-class hub for AI and ensure that AI is human-centred and trustworthy*". Resulting of the recently proposed AI Act, a strategy is suggested in which there are legislative instruments together with codes of conduct for the less "*risky*" AI algorithms in order to control and regulate the use of AI in the European Union [EC, 2021a]. When focusing on levels of automation in decision-making processes it is interesting to know whether there are European rules and or guidelines on automated decision-making. The GDPR states that "*the data subject shall have the right not to be subject to a decision based solely on automated processing*" (Art. 22.1, GDPR). Thus full automation shall not be possible unless it is allowed so by (national) law still protecting the person's individual rights. Some regulations on automation via artificial intelligence in the public domain will be mentioned in the next sections in order to have a brief understanding of what is already in place and what the struggles are. First, some regulations from European, National and a more local level will be discussed before briefly elaborating on the role of civil servants and local governments such as municipalities

2.3.1 Artificial Municipalities

Municipalities and cities have been using AI in multiple instances in the Netherlands, though it may be in a pilot form or in a test phase. Examples of this are for example the automated parking control algorithm or an algorithm aiding in tracking illegal renting out properties as holiday homes [Cath and Jansen, 2023]. As Cath and Jansen [2023] also mention, the Dutch government has released *"Het Algoritmeregister"*, registering every algorithm that is used by a public organisation within the Dutch government aiding in providing service to the public [Algoritmeregister, 2020]. Research has been done by the Association of Netherlands Municipalities (Vereniging Nederlandse Gemeenten (VNG)) on the current and future laws and regulations on the use of AI in the public domain and specifically in municipalities [VNG, 2020a]. Three levels where these developments take place have been discussed and are summarised in the table below.

As one can conclude from table 2.1, more guidelines are defined for the use of AI in the public domain, but not specifically on local governments and executing governmental institutions. And from a more local perspective, it is harder to make the laws and regulations from a European perspective more concrete on their own

Table 2.1: Laws and Regulations for the Use of AI by Municipalities VNG [2020a]

Levels of Laws and Regulation	Key Points	Goal
European Commission	• Restrictions by law for 'high-risk' AI appliances • Clarifying the process of accountability • Obligated conformity assessment • Voluntary quality mark for 'low risk' AI appliances • European governance on international collaborations	Ensuring safe, human-centred, meaningful and trustworthy AI, where the European Union is the leading hub for AI.
National Guidelines	• Risk awareness • Explainability • Data Recognition • Auditability • Responsibility • Validation • Testability	Providing guidelines rather than laws on the development and use of algorithms by the government for the purpose of informing the public.
Local Principles by Municipalities	• On the digitalisation aspects of the public space.	Aims to impact the data gathering by AI and the way governments operate with AI.

appliances. Mikalef et al. [2022] found that local governments do prefer governmental pressure on the development of AI deployment in particular. They stated that municipalities are very eager to take over rules and guidelines from a national level on this topic in order to align themselves with a national or even European strategy. The VNG, therefore, came with a call for attention and with some advice on the practicality of the current AI regulations [VNG, 2020b]. This research hopes to provide a wide and flexible advice implementable for all municipalities and more practicality and attainability in the operationalisation phase when civil servants will have to engage with the AI in decision-making on different levels of automation, which will help to accelerate the deployment of safe, responsible and meaningful AI in local governments [Mikalef et al., 2022; Siebert et al., 2022].

2.3.2 Digital Civil Servants

UNESCO presented a report stating how civil servants should adjust in a digital transformation of governmental institutions [Vinadio et al., 2022]. They stated that 5 aspects are important for civil servants: *trust, creativity, adaptability, curiosity and experimentation* in order to effectively transform into a digitised government, for which altruistic leadership would be very beneficial. Next to this, they state the importance of knowledge and know-how for civil servants in order to work with technologies that help in digitising the public domain. This also comes back to the deskilling that occurs when automating tasks of civil servants. *reskilling* or *upskilling* would be a solution for a better implementation of AI [Rafner et al., 2021]. Different applications such as virtual assistants, robotic process automation and analytical prediction are ways of overtaking tasks of civil servants, but clear-cut laws on a European or national level are missing. The aim of the European Commission is to provide laws and regulations at the end of 2023, after which the Dutch government has already made clear to follow swiftly with implementable laws and regulations on a more practical level [Leeuw, 2022]. This research aims in providing insights into the current situation of civil servants in a digitising government and how experts view the role of civil servants when implementing new technologies such as AI.

2.3.3 Artificial Local Governments

There is a distinction between local governments that have been eager to explore the new possibilities through the implementation of AI and local governments that have preferred the wait-and-see method [Yigitcanlar et al., 2022]. Opportunities, challenges and risks arise with the vast implementation of new technologies where clear laws and regulations are lacking [Yigitcanlar et al., 2022; Roehl, 2022]. Up to now, most regulations and guidelines exist on the type of AI, which should be safe, transparent and fair. Whereas these guidelines are lacking in advising on the level of automation and on the implementation of AI where it is working together with the human.

Another interesting aspect of using artificial intelligence in local governments is to be able to say that throughout the Netherlands these different local governments should deliver the same type and same quality of service when they use the same kind of AI to assist at the same level of automation to the civil servant in the decision-making process. This would be in line with one of the fundamental rights of citizens where they receive access to equal service [Simonofski et al., 2022]. Nonetheless, Simonofski et al. [2022] also elaborate on the right of *human* public service, with which full automation of decision-making over individual citizens is excluded. This again also brings forward the notion of meaningful human control. You have to make sure that enough "human" is present in providing the public service to still adhere to this fundamental right [Siebert et al., 2022; Methnani et al., 2021; Simonofski et al., 2022]. The notion of meaningful human control will be elaborated on later on in this report as this has been a very popular concept in late literature on the issues of AI implementation. This research dives into the trade-off experts make between human control and automation benefits where these legal frameworks need not be disregarded as they will be developing swiftly over time (AI ACT).

2.4 CONCLUSION

Methnani et al. [2021] mentions the need for research on when, to who, and how to transfer control. This paper tries to aid in this purpose by diving into the trade-offs made when implementing AI in the public domain, specifically focusing on the role of civil servants in local governments. The level of automation determines to what extent the human and the AI will cooperate in the decision-making process.

Where previous research has mainly focused on the technological aspect and on how to determine what type of AI would be sufficient and how can AI make decision-making more efficient, this research strives to focus on the combination of technology and expertise while taking into account the societal impact that this hybrid human-AI cooperation has [Engin and Treleaven, 2019; Bullock, 2019; Araujo et al., 2020; Kuziemski and Misuraca, 2020a]. There is still a large need for more empirical knowledge on the implementation of AI in hybrid forms with different levels of automation [König and Wenzelburger, 2021]. especially in the public domain and regarding all the stakeholders present. This research aims to provide more insights into the current knowledge gaps of the perspectives of experts focusing on the role of civil servants and local governments, using different levels of automation in their decision-making process to have a certain form of cooperation between civil servants and AI. Due to the developing laws and regulations for (local) governments, experimenting with AI has been a challenge, and this research aims to provide some tools and handlebars for the responsible implementation of AI. The goal is to add knowledge to the question of how to responsibly implement AI in the public domain and to provide guidance to those willing to experiment with new technologies in supporting civil servants and their developing role in new forms of human and technology collaboration.

The role of civil servants is crucial in decision-making processes, and their discretion and human control need to be considered when determining the appropriate level of automation. The review also discusses European guidelines and regulations concerning AI in the public domain, emphasizing the need for human-centred and trustworthy AI. Furthermore, it touches upon the evolving role of civil servants in a digitized government and the importance of their knowledge and adaptability. The research aims to provide insights into the perceptions of civil servants and their role in implementing AI, ultimately contributing to the responsible and meaningful use of AI in local governments.

From the literature, we can conclude that human control is of high importance in controlling and implementing a safe system. Human control is mentioned in multiple different concepts such as human-in/out-of-the-loop, meaningful human control, etc. It can therefore be concluded that looking at the civil servant as the human controller and operator is of key importance and needs extra investigation.

Based on this literature review 4 observations have been made that are listed below and combine the knowledge gained from the literature research into the topic and knowledge gap for this research. These observations will be tested in this research to see whether the domain experts also consider these observations as important and how this research could help to build upon these observations and eventually provide guidelines on implementing a responsible level of automation. The next chapter will explain the methodology of this research and how this research aims to answer the research question and thus provide guidance for those willing to experiment with AI in the public domain. The research approach for every sub-research question will be elaborated on.

O1: Human autonomy and quality of service are the main criteria that determine the level of automation in automating decision-making by civil servants in the public domain in the Netherlands.

O2: Cooperation between the scientific, public and private domains is needed to identify the criteria and responsibly implement AI in hybrid human-AI systems with meaningful human control.

O3: AI can be used to improve the civil servant's discretion for different decision-making processes, which can contribute to a higher quality of decision-making.

O4: Researchers acknowledge a new role of civil servants where human autonomy is more important for responsible AI implementation.

3 METHODOLOGY

In this chapter, the research approach is presented and explained based on the findings from the literature that resulted from the previous chapter. After defining the research scope, the strategy that is chosen to answer each sub-question is explained. Next, the resulting case study methodology is presented followed by an extensive presentation on the data collection. This includes information gathering and data analysis through semi-structured interviews and validation of the gathered input from the interviews through a survey.

3.1 RESEARCH APPROACH

The thesis is focused on the implementation of AI in the public domain and more specifically on how certain criteria and their trade-offs can determine a responsible level of automation for different decision-making processes, specifically focusing on the role of the civil servant. Aiming to answer the following research question: *What trade-offs are made between efficiency and human control when determining a responsible level of automation for different decision-making processes in local governments?* This research tries to understand what impact digitalisation has on the public domain and how public institutions can adjust to and adopt new strategies in order to fulfil all their citizens' needs. The research has an exploratory approach where understanding is gained from an in-depth analysis of different ways of implementing AI in a governmental organisation working in hybrid human-AI systems, analysing, discussing and reviewing the different criteria.

In order to answer the research question, a qualitative method will be used to analyse the complexities that arise when trying to address the impact of decision trade-offs and criteria for the implementation strategies on the level of automation of decision-making tasks in public organisations. This will focus on defining the exact trade-offs that take place and how public organisations see this for themselves. A more quantitative method will be used in order to try to gain insights into the trade-offs and decisions that experts from the three domains make when deciding upon the automation level of AI for a certain task in the public domain. This is done by looking at 2 different implementation scenarios for 2 different cases, resulting in four scenarios in total, by means of a vignette experiment. This will be elaborated on later and provides more insights into the trade-offs that could impact the preferred and most responsible level of automation. Below the research scope will be defined before diving into the different research methods that are used to answer the different sub-questions.

3.1.1 Research Scope

The public domain in the Netherlands is, relative to other (European) countries, highly digitised and is, therefore, an interesting case to explore the criteria for implementing AI and the level of automation that is most beneficial [UN, 2020; Delfos et al., 2022; Dieter O., 2022; NordicHQ, 2023]. Within the public domain, this research will focus on local governments and executing governmental organisations where clear guidelines on the use of AI are lacking or unspecified. Both the options of AI algorithms from within the government or from an external private company are considered, though the main focus lies on the latter where there is a collaboration between private and public organisations. Since the end of 2022, the Dutch government has already made another improvement in the implementation and transparency of AI by launching the algorithm register [Rijksoverheid, 2022a]. More and more research is being done on the implementation of AI, the different types of automation tools and the opportunities and challenges [Delfos et al., 2022]. But no research has focused on how to determine a responsible level of automation that an AI can or should fulfil in a public organisation and what the criteria are to make such a decision. How can these trade-offs be made insightful and comprehensible for policymakers, civil servants and citizens in the public domain, in order to responsibly implement AI working together with the civil servant? It makes sense to focus on the civil servant since they will be the operator of the AI in the strategy chosen by policymakers, and could be considered the controller of the hybrid human-AI system.

Control is vital and a key aspect in this topic since in Dutch governmental AI implementations, plenty of errors have occurred in the past where the citizen or civil servant perspective was not or not extensively included in the policy-making process. This may sound contradictory since the public interest should be the key driver for governmental institutions. Though it became clear through for example the infamous 'Toeslagenaffaire', that the citizen perspective and their wishes had gotten out of sight. A more citizen-centred decision-making process should become the norm. Therefore it is important to include all stakeholders' perspectives in deciding upon the level of automation of even the smallest administrative tasks and how this differs per tasks by governmental institutions Luk [2009]. The perspectives of laws and regulations will be left out of this research scope though as mentioned before it is very important to be aware of those. They are however still changing and being updated on a monthly basis which makes it hard to take into account for now. This research, therefore, focuses on the needs and role of the civil servant from the perspective of experts from the public, private and scientific domains. Also focusing on how civil servants can be supported as responsibly as possible by these new technologies such as AI. Including European and national laws and regulations in future research will result in a more complete picture with hopefully clear, refreshing and helpful insights for policymakers. For this research, it is too big to tackle the issue of the whole Dutch government and therefore this has been scoped down to local governments, eventually, resulting in policy advice on how AI could be implemented responsibly to benefit civil servants and thus the public service.

3.1.2 Strategy per Sub-Question

This section explains the methodological strategy that will be used in order to answer each sub-question. This will result in widely and extensively gained insights due to the different strategies. As mentioned before, this research can be split into 3 research blocks. In figure 3.1 you can see how these sub-questions align with the different research blocks and what research methods will be used to answer these questions. The three sub-research questions also aim to test the observations found in the literature from chapter 2 which will be mentioned again after the strategy per sub-question.

SQ1: *What criteria do experts see in automating direct decisions making for citizens, regarding different context factors?*
In order to answer the first sub-question, desk research will be done. A desk research gives all the criteria and requirements that have been deemed important in scientific and grey literature for decision-making for implementation of decision support tools such as AI, Machine Learning (ML), etc. With this

question, the first observation can already be tested partially by checking the wanted outcomes for these criteria. This desk research also results in a number of frameworks and theories that have been used in previous literature in the process of determining a level of automation for decision-making processes. Semi-structured interviews will validate the findings from the desk research and add on the results by providing a selection of the main criteria according to experts from the three domains that will be interviewed (private, public and scientific).

SQ2: *How do experts make the trade-off between human control and efficiency gain in deciding on a level of automation by looking at different context factors resulting from the interviews?*
The second sub-question will be answered through semi-structured interviews and using the results and frameworks found during the desk research. Also focusing on the known trade-offs and how experts from these three domains see these trade-offs and how they would envision responsible use of AI. Especially regarding the level of human control by the civil servant and the role of the civil servant. The aim is to be able to advise on a certain level of automation, which will be elaborated on in the next chapter. Different trade-offs may result in a different level of automation, especially regarding the context and type of a decision that is to be supported by an algorithm. And this is exactly where discretion by civil servants has been considered so important [Mitrou et al., 2021]. As Mitrou et al. [2021] mention, the degree of discretion and human control are needed to determine the level of impact that an algorithm may have in decision-making processes in public organisations to provide a responsible implementation.

SQ3: *How do the trade-offs help in determining a responsible level of automation for different decisions, regarding sensitivity and impact on citizens?*
Based on the results from the interviews and from the literature review, a survey will be conducted with people working in the public, private or scientific domain and working with AI. With a focus on civil servants, they are asked how they value certain criteria in a certain level of automation for a specific task. Asking them how willing they would be to make use of such a system where AI and humans collaborate in the decision-making process that directly impacts citizens. Based on this survey the main criteria resulting from the interviews can be tested for different cases where the context factors also differ. The final observation can be tested after answering this question, trying to find what context and what situation would impact trade-off and thus the decision for a certain level of automation.

Observations:

O1: Human autonomy and quality of service are the main criteria that determine the level of automation in automating decision-making by civil servants in the public domain in the Netherlands.

O2: Cooperation between the scientific, public and private domains is needed to identify the criteria and responsibly implement AI in hybrid human-AI systems with meaningful human control.

O3: AI can be used to improve the civil servant's discretion for different decision-making processes, which can contribute to a higher quality of decision-making.

O4: Researchers acknowledge a new role of civil servants where human autonomy is more important for responsible AI implementation.

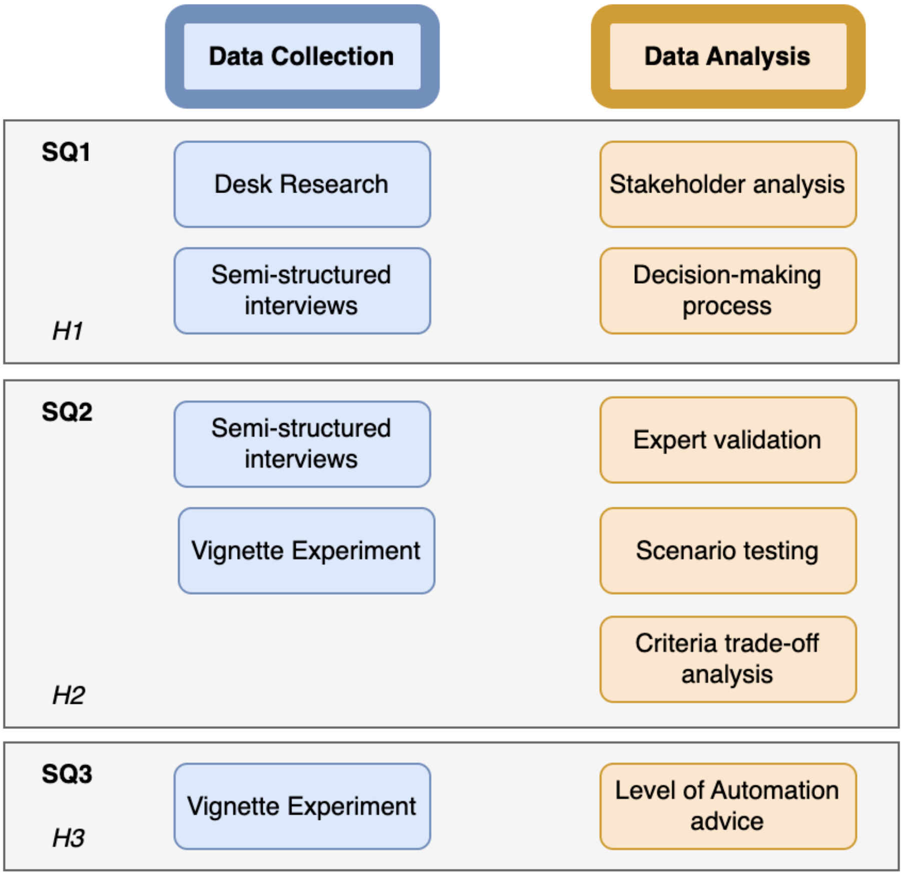


Figure 3.1: Research Strategy per Sub-Research Question

3.2 CASE STUDY

For the survey, a case study will be used to give an in-depth analysis and explanation of the problem at hand. Yin [2014] defines a case study as *"An empirical inquiry about a contemporary phenomenon (e.g., a "case"), set within its real-world context, especially when the boundaries between phenomenon and context are not clearly evident"*. Where a case study does not only look at isolated instances but takes them to a broader perspective and extends further than just exploratory purposes. Yin [2014] discusses 3 steps in designing a case study: *"Defining a "case", Selecting one of four types of case study designs, and Using theory in design work"*.

In this thesis, a holistic case study is chosen for a case study within the public domain focusing on a decision-making task where a sensitive and a less sensitive decision are being made for individual citizens by civil servants. This is defined by Yin [2014] as a *"single case design"* with a *"multiple units of analysis"*. This case study focuses on civil servants working on decision-making for individual citizens in the Netherlands which makes for a geographically specific case study on a certain organisational level. The criteria for this decision and how civil servants come to a decision for these types of tasks will be explained later on before diving into the trade-offs civil servants would make when this task would be automated to a certain level. The two different scenarios that will be looked at have different values for certain criteria that are the result of the desk research and the interviews also looking at different context factors. This enables easy comparison between different cases regarding their context factors and also in comparing a case by looking at the different criteria that resulted from the interviews. This survey aims through the usage of two real-life cases to identify trade-offs that experts make when determining a responsible level of automation.

3.2.1 Case Selection

As explained in the previous chapter it has become clear that European or National guidelines are getting more prominent and present these days. However, how to concretely implement and follow them is a challenge for local governments. This thesis focuses on smaller local governments in the Netherlands, such as small municipalities. The aim of this research is to provide advice and insights into automating decision-making by civil servants that directly impact individual citizens. Irresponsible implementation of AI could result here in a new 'Toeslagenaffaire' where data is wrongfully used, where biased data is used, where no ripple and side effects are considered, where irrelevant criteria are misused, etc. Councyl also has an interest in investigating the public domain as a possible market for AI implementation but mainly wants to provide insights into the public domain and how and where to responsibly implement AI so that it benefits the organisation and the citizens. Within these boundaries, different context factors will be analysed to see what their impact is on a responsible implementation of AI in the public domain. These 2 cases that will be compared will be elaborated on when the vignette experiment, where the two cases will be compared, is explained.

3.3 DATA COLLECTION

As mentioned earlier, the data collection will contain literature research, semi-structured interviews and a survey. This is needed to answer the sub-questions and finally answer the main research question. This section will elaborate on these 3 methods of data collection and how these will be formed.

3.3.1 Desk Research

Desk research is needed for a greater understanding of the different kinds of classifications that have been made on the level of automation in scientific literature. From the literature review, the concepts and criteria around the implementation of AI in the public domain, that have been discussed in scientific research, will be found and analysed. Clear demarcation on the type of criteria that are necessary for this research will also be made by diving into the other concepts around value, requirements, etc. Although for some people AI may sound like something relatively new and futuristic, it has already been around since 1950 when Turing spoke of computing intelligence [Turing, 1950]. AI is a technology that gained a massive boost in the last few years and especially in the last few months due to the ChatGPT, so plenty of attention has been given to this subject and enough papers have been written on its potential, challenges and pitfalls. Due to the hype of late, plenty has been written on "new" automating technologies, not only on AI but also on the different types of ML, Neural Networks (NN) and AI, regarding the automation of human decision-making tasks by machines [Siau and Kam, 2006; Meinert et al., 2019; Silveira, 2005; Simonofski et al., 2022; Asadi et al., 2014; Hood, 1991]. Existing frameworks and notions on how to implement these technologies responsibly will be analysed to construct an understanding of how AI could be safely and responsibly implemented in the public domain. These findings will be used to shape the next steps of the research and to be able to dive into the criteria and context factors of public-specific decision-making processes by civil servants, using a certain level of automation that results in hybrid human-AI decision-making.

3.3.2 Semi-Structured Interviews

The semi-structured interviews will gather data on the same topics as the literature review though with a more specific focus on the implementation of AI in the public domain and on the perspectives of three different domains, the private, public and scientific domains, regarding a set of criteria and the contextual factors that play a role in determining a responsible level of automation for the decision-making. Due to its semi-structured setup, there will be room for the interviewees to interpret the questions freely and come up with novel ideas and concepts that could add value to the data found in the literature research, "following the respondent's train of thought" [Bolderston, 2012]. The questions need to be stated clearly beforehand but may be rephrased when asked to the interviewee based on the needs of that interview. A lot of room should be created to be able to let the interviewee decide what they want to see within the boundaries that you have set. And these boundaries may be flexible if the topics discussed are thought to be valuable for the research. Though all questions will have to be asked to all respondents. The interviews will also be used to see if those criteria found in the literature align with the real-world practice of implementing AI in the public domain. For a semi-structured interview, it is important to select key persons for the case study that you are going to dive into.

In this thesis, it would be interesting and important to have a look at those involved with implementing AI in the public domain or even more specifically, in the decision-making that is being done in local governments. The responsible implementation of AI should be key and an important part of this is the user or cooperating human and whether this person understands their role and the usage of AI. These interviews will help to answer research sub-questions 1 and mainly 2 since this sub-question requires insights from experts working with implementing AI or in the public domain. Insights from these 3 domains provide a valuable and complete overview which can be used to look at the trade-offs of some main criteria, the impact of contextual factors on the level of automation and on the level of control by the civil servant and how the role of the civil servant could change.

1. *Public*: 3 experts were interviewed on their viewpoint on the role of civil servants and on the criteria that they think are important or should be present in determining the suitable level of automation.
2. *Private*: 3 experts were interviewed on their experience with implementing AI applications in the public domain and how civil servants see these innovations and how they deal with it. Trying to get more insights into how civil servants fulfil their role as human controllers of AI automation systems.
3. *Scientific*: 3 experts were interviewed on their perception towards responsible AI implementation in the public domain and how the civil servant should fulfil the role of meaningful human control in automated decision-making processes in the public domain.

All experts were also asked to elaborate on the possibilities and challenges that they see when discussing the implementation of different levels of automation in hybrid human-AI systems in decision-making processes. Especially regarding different context factors and different phases In the decision-making process and roles of the civil servant. Recommendations can be made on a number of levels as presented in table 6.1 at the end of this chapter.

Table 3.1 below shows the experts that had been included in this research for the interviews. All experts cover a different domain and different expertise though, for each domain (private, public and scientific), three experts have been interviewed in order to get as much information as possible. Though more experts may result in more information, three experts for the three domains were considered to be sufficient beforehand and also turned out to be sufficient as most of the questions were satisfied in their answers after 8/9 interviews. These interviewees were mainly gathered through the concept of snowball sampling and by visiting congresses and conferences where people with expertise on civil servant decision-making or AI implementation in the public domain. Snowball sampling entails asking the interviewees to aid the researcher in finding other suited subjects for the interview [Goodman, 1961]. This has been a famous method for recruiting respondents as potential subjects in qualitative research. These subjects did have to comply with a level of expertise on the implementation of AI in the decision-making process by civil servants, from the perspective of their domain.

Table 3.1: Interviewees

Interviewee	Domain	Role/Function
Interviewee 1	Science	PhD Researcher TU Delft
Interviewee 2	Science	Senior Researcher TU Delft
Interviewee 3	Public	Coordinating Policy Officer in the domain of AI
Interviewee 4	Private	Software Develop & Data Analyst
Interviewee 5	Private	Deep-Fake Expert
Interviewee 6	Private	Project-Manager
Interviewee 7	Science	Researcher in Informatics and Responsible Implementation of Technologies
Interviewee 8	Public	Policy Advisor on Ethics
Interviewee 9	Public	Project-Manager

3.3.3 Survey Analysis

A survey experiment will be held in order to determine the main trade-off between a most frequently discussed criteria, resulting from the interviews, and how this relates to the level of automation. This survey experiment focuses on answering the third research question as well as providing insights to answer the main research question together with the gathered information in the first two sub-questions. A vignette experiment will be used as a way of conducting this survey [Atzmüller and Steiner, 2010]. A vignette is a short and brief description of a person, object or situation where certain variables (factors) can have different values [Atzmüller and Steiner, 2010; Silva et al., 2019]. A vignette experiment is special in that it allows to combine of several explanatory and context factors which creates a way more realistic scenario than in other forms of survey [Atzmüller and Steiner, 2010]. As Atzmüller and Steiner [2010] explain, the flexibility of presentation of these scenarios, where multiple criteria and context factors are combined, also aids in the realism of this research method. In this thesis, two cases will be provided where they both differ in the most important contextual factors that can be concluded from the desk research and interviews. Each case will have two scenarios with each scenario having a certain trade-off present between the main criteria, resulting in four scenarios in total. The vignette experiment also aims to get an understanding of people working in the three domains as mentioned in the interviews as well, public, private and scientific. The respondent is asked to grade each scenario on whether they would be willing to use or implement such a system. By providing different scenarios for the same case, with different values for a few criteria, the expert considers different levels of automation for the same sensitive decision-making case, where a decision is made for an individual citizen.

Context factors are added to provide more detail on the case so that the civil servants will also better understand the scenario and leave as little room for their own interpretation as possible [Atzmüller and Steiner, 2010; Sheringham et al., 2021].

The survey will also include an optional section where respondents can explain why they grade certain scenarios differently with the goal to get some more insights into the reasons why they would be willing to use a certain form of hybrid human-AI decision-making. As Kuziemski and Misuraca [2020a] already mentioned, there is a need for survey-based experiments to test citizens' acceptance but also to understand the way civil servants look at the automation of their tasks. In order for citizens to accept the automation of governmental decision-making the role of civil servants, and whether they also do agree with the way it is implemented and the level of automation, is very important. Therefore this research tries to find the most responsible way of implementing AI, by using insights from experts, in such a way that it can support the civil servant as much as possible. The respondents for this vignette experiment were identified at a conference on the responsible implementation of AI in the Dutch public domain. This was done to get respondents that already have prior knowledge on this topic or may even have working experience from their domain. Most of the subjects present work in the public domain which is useful as this research aims to say something on the role of the civil servant. Though by including some subjects from the scientific and private domains, a comparison between these domains and more complete picture can be made. Future research will be needed on their acceptance and that of citizens. Defining a responsible level of automation with a sufficient level of control by the civil servant is however first needed. The subjects of the public domain will give valuable insights into the way civil servants think and operate in order to already give some advice on this. The analysis through a vignette experiment will be elaborated on more, later in this chapter.

3.4 DATA ANALYSIS

This section elaborates more on how the data retrieved from the above methodologies will be analysed. The different steps will be discussed here before concluding on this chapter and following with the more in depth desk research into different levels of automation in the next chapter.

3.4.1 Stakeholder Analysis

Relevant stakeholders need to be identified and analysed in order to understand what experts need to be consulted for the responsible implementation of AI in public decision-making in smaller municipalities. As already mentioned in chapter 2, multiple stakeholders are present when implementing AI responsibly in the public domain, making this a complex multi-actor system. Identifying the stakeholders is an iterative process though with a clear demarcation of the research topic and the research phenomenon, it becomes easier to identify the most important and relevant stakeholders [Reed et al., 2009]. The most relevant stakeholders should all be included in the decision-making processes and design phases when deciding upon a certain level of automation. There is a risk of overlooking important or relevant stakeholders when trying to identify them before diving further into the topic, though for this research a clear demarcation has been made based on some criteria stated by the researcher.

As mentioned earlier, direct and indirect stakeholders can be identified where the direct stakeholders are directly involved with the new technology, policy or the implementation of those. Indirect stakeholders will not have a direct impact on this decision-making but will be impacted by any decision made on the implementation of new technology. Stakeholders will be analysed by looking at the boundaries of the research domain and the case study. Following Kool et al. [2010] and Makoza

[2019] stakeholders need to be identified before they can be analysed. The identification of stakeholders in ICT policy implementation can be done in multiple ways looking at the policy goal, at the responsible parties or at for example the values that come about and how these are impacted [Makoza, 2019]. The latter is interesting also from a VSD point of view, where the values of different stakeholders are mapped when considering AI implementation in any form of public decision-making. Stakeholders can be identified where their values are either impacted by a policy implementation or when they can impact other stakeholders' values. Reed et al. [2009] propose a categorisation of stakeholders where they can be classified according to their interests and the influence they have on the problem. For this research, the stakeholders will be limited to the case study with the geographical demarcation of the Netherlands and also looking at smaller governmental organisations that operate on a more local level rather than a national level. Stakeholders responsible for AI creation or implementation will be included together with those having to work with the AI implementation. These stakeholders are prioritised to be involved in the decision-making process and can contribute to a responsible level of automation of public decision-making on individual citizens.

3.4.2 Interview Data Analysis

The data gathered via the interviews needs to be analysed in a way that can be used for answering the first and second research questions. The first research question will mainly help in supporting the findings from the desk research and adding to this knowledge where needed and possible. For the second research question, the interviews will have to provide all the needed data combined with the data from the desk research as a knowledge foundation. The interview is structured in three parts asking about the criteria that could impact the responsible level of automation, the context factors that play a role and the changing role of the civil servants. The interview follows the ethical guidelines by Bolderston [2012] including consent forms, privacy and confidentiality considerations. Open questions will give the interviewees the possibilities to give different answers and elaborate on them. The interview is recorded when the interviewee has agreed with this. The entire interview will be transcribed using Microsoft Word, where interesting quotes will be literally quoted when given permission by the interviewee. Following McLellan et al. [2003], the transcript should contain what is necessary for the research. For this research, the transcript does not entail the entire interview but is filtered into the aspects that give answers to the question and that provide new insights. If the interviewees mentioned other topics that were considered informative or useful to use and mention, these have been included as well. The transcriptions of the interviewees will be coded in order to understand the three parts better, trying to gain insights into the reasons for automating decision-making in the public domain according to these experts. But also understanding the contextual factors that play a role and what the role of the civil servant should be in order to accomplish a responsible AI implementation.

3.4.3 Vignette Experiment

As mentioned, the vignette experiment provides a unique method in which multiple context factors and criteria can be tested, also by looking at the trade-offs that people make between these criteria and context factors [Atzmüller and Steiner, 2010]. Different types of vignette experiments exist but this one focuses on the main group of criteria and context factors to be able to determine the smaller interactions and trade-offs between these criteria and context factors. Where Atzmüller and Steiner [2010] mention, three different types of vignette experiment designs exist where this research focuses on the *within-subjects design*, where every respondent gets the exact same vignettes to judge. This is beneficial due to the small amount of vignettes and limited number of respondents. The construction of vignettes has four requirements as stated by Heverly et al. [1984]:

1. The constructs of interest that are needed to create the vignettes need to be identified and defined. In this research, these are gathered from a literature review and ranked and validated through interviews.

2. Other components and context factors need to be examined and created for the vignettes. This research uses expert interviews to gain insights into the contexts of certain scenarios.

3. The validation of the vignette components is needed to get to a final set of vignettes that are of interest to this research. This is done by combining the literature review with the expert interviews and by going back to the main research question to determine whether this could help in advising on a responsible level of automation.

4. The last step is the construction of the vignettes from the components mentioned above. Due to the smaller group of respondents and the small number of components that are present in the vignettes, the components are not randomly distributed.

This set of vignettes that is presented to the respondents is a subset of the entire vignette population [Atzmüller and Steiner, 2010]. From the data gathered from the interviews and the desk research, multiple scenarios or vignettes could be created by including all the mentioned criteria and context factors. However, to be able to get concrete results on the trade-offs between the criteria and context factors that are believed to be the most important, a subset of vignettes is created. As mentioned by Evans et al. [2015] and Langer [2016] the ideal set of vignettes should be a summary that clearly describes realistic scenarios, and can be fictional, where the research variables and context factors variate. This means that not all the criteria and context factors are taken into account for this vignette experiment although it is advised to conduct further research into these other criteria and context factors, especially in different scenarios and different decision-making processes. Evans et al. [2015] created a list of recommendations for the content of vignettes and how to clearly present them to the respondents, this can be seen in table 3.2. By looking at detailed vignettes, with a sub-set of the identified criteria, one can focus more on the reasons behind certain choices of the respondents which are in line with the exploratory and qualitative nature of this research [Sheringham et al., 2021]. This is made possible also due to an open question at the end of each vignette where participants are asked to give their reasoning behind their choice. This open question is voluntary.

Following the 3 steps mentioned by Aguinis and Bradley [2014], combined with the requirements mentioned above, a clear and sufficient vignette experiment can be created for exploratory and qualitative analysis of different levels of automation in the public domain following the criteria and context factors that result from the desk research and interviews. These three steps can be seen in figure 3.2, within this figure they mention 10 decision points. Since most of these have already been mentioned by looking at the requirements by Heverly et al. [1984] and the recommendations by Evans et al. [2015], these will not be extensively discussed. These three steps in this figure are however clearly represented and will be followed in chapter 5, where the vignette experiment is created and the results are presented. This report will advise on how to look at the criteria and context factors that are considered to be important by the experts in the domain, and how to determine a responsible level of automation for these specific tasks following the results from

Table 3.2: Recommendations for vignette content, adapted from Evans et al. [2015]

	Vignettes should
1.	Derive from the literature and/or clinical experience
2.	Be clear, well-written, and carefully edited
3.	Not be longer than necessary (typically between 50 and 500 words)
4.	Follow a narrative, story-like progression
5.	Follow a similar structure and style for all vignettes in the study
6.	Use present tense (past tense only for history and background information)
7.	Avoid Placing the participant ”in the vignette” (e.g. as first- or third-person character)
8.	Balance gender and age across vignettes
9.	Be as neutral as possible with respect to cultural and socio-economic factors
10.	Resemble real people, not a personification of a list of symptoms or behaviours
11.	Be relatable, relevant, and plausible to participants
12.	Avoid ”red herrings”, misleading details, and bizarre content
13.	Highlight the key variables of interest, facilitation experimental effects
14.	Facilitate participant engagement and thinking by including vague or ambiguous elements
15.	Cover all pertinent variables (or omit selected variables for specific purposes)

the vignette experiment. The exact cases and scenarios that will represent the vignettes will be explained in the chapter 5 where the information gathered in the desk research and the interviews are combined.

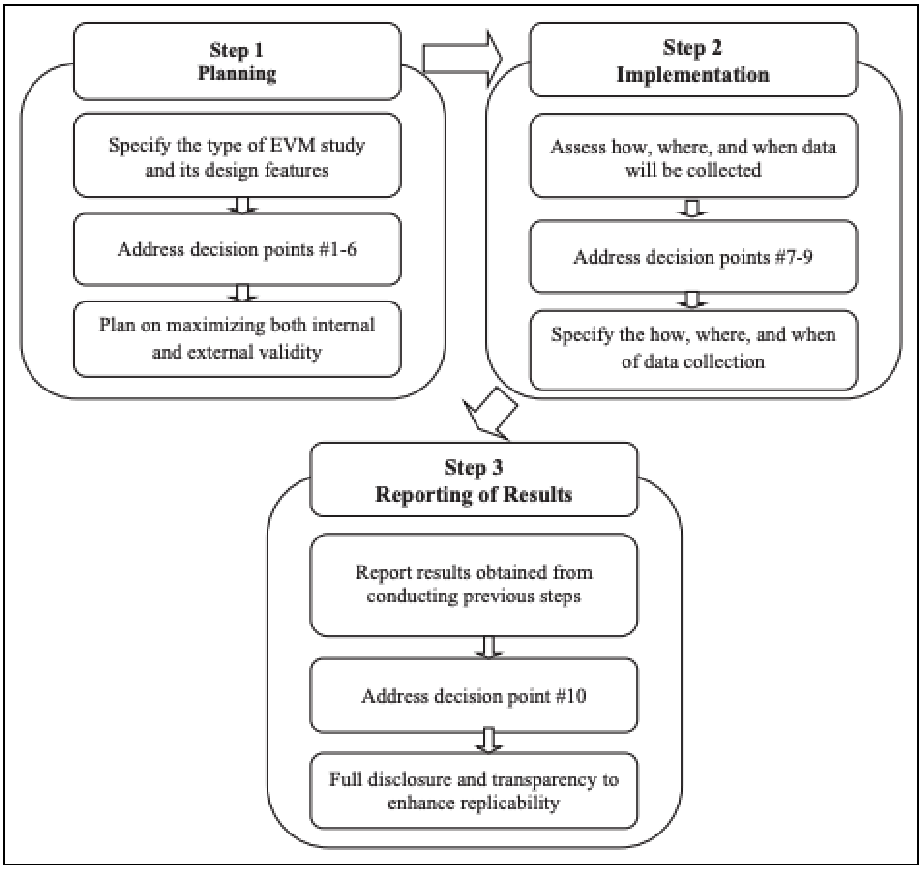

Step 1
Planning
Specify the type of EVM study and its design features
Address decision points #1-6
Plan on maximizing both internal and external validity
Step 2
Implementation
Assess how, where, and when data will be collected
Address decision points #7-9
Specify the how, where, and when of data collection
Step 3
Reporting of Results
Report results obtained from conducting previous steps
Address decision point #10
Full disclosure and transparency to enhance replicability

Figure 3.2: Summary of steps and decision points in conducting an experimental vignette methodology study, adapted from Aguinis and Bradley [2014]

3.5 CONCLUSION

This chapter first focused on the research scope and on the different research questions before briefly elaborating on the case study. The different data collection and analysis methods and techniques for each sub-question have been elaborated on. Desk research is conducted for stakeholder analysis and for criteria and context factor identification. This will be elaborated on in chapter 4 defining the different levels of automation that have been mentioned in the existing literature, together with some frameworks that also (partially) look at the criteria and trade-offs. These frameworks will be analysed through desk research and by looking at the trade-offs between these criteria, together with frameworks from the literature, which will constitute the foundation for testing the criteria and context factors examined in the interviews and surveys. Interviews and a vignette experiment will be used to test the main criteria and context factors and to identify and better understand the respondents' trade-offs when deciding upon a responsible level of automation. Focusing on the role of the civil servant in hybrid human-AI decision-making. The results and analysis from the interviews and survey validation will be presented and explained in chapter 5.

The interviews try to validate whether human control by civil servants is indeed the main aspect of implementing a responsible level of automation. Also aiming to identify the main criteria, context factors and insights into the role of the civil servant and lacking knowledge of policymakers willing to experiment with this. In the vignette experiment, a comparison will be made between two cases of decision-making in the public domain by smaller municipalities that differ in certain context factors resulting from the interviews. Two scenarios or vignettes will be created for each case where respondents are asked how willing they would be to make use of such a hybrid human-AI system. The aim is to be able to determine, or at least provide guidance in determining, a responsible level of automation for a particular decision-making process based on these trade-offs. These different levels of automation and the criteria are identified and explained in the next chapter.

4 LEVEL OF AUTOMATION

This chapter aims to provide insights into the different levels of automation that have been discussed in the existing literature and what implementations have been used so far in practice. This research aims to find how the trade-offs between the main criteria and context factors determine a responsible level of automation when automating a certain decision task, following the methods and techniques from the previous chapter. It is therefore insightful to first dive into the existing classifications and on the criteria that have already been identified before validating and testing these during the interviews as will be elaborated on in chapter 5. This chapter focuses on conceptualising and operationalising the level of automation.

After diving into the existing classifications, the different criteria that establish these different levels of automation will be elaborated on before touching upon some trade-offs and contextual factors. Next, the levels of automation will be discussed according to two different existing models where different stages could be automated as discussed by Parasuraman et al. [2000] and where different roles that are present in a decision-making process are discussed by Cummings and Bruni [2009] using their Human-Automation Collaboration Taxonomy (HACT) Framework.

As [Rombach and Steffens, 2009] mentioned, the implementation of a new (information) technology in the public domain asks for "four dimensions of action": *strategy, process and organisation, technology* and *project and change management*. This chapter tries to make a first step towards these four domains, aiming to provide insights that can help in defining a well-suiting strategy. Involving respondents and subjects from the three involved domains and thus aiming to include process and organisational aspects while also looking at the technology aspects of the level of automation. Altogether trying to aid in providing handlebars and guidelines in the process of digitising the public domain and governmental tasks, with a specific focus on the role of the civil servant.

4.1 EXISTING CLASSIFICATIONS

In order to determine the level of automation that AI may have on human decisions in administrative tasks in the public domain different aspects of these tasks need to be discussed. The type of task that the AI algorithm will support is for example important. To give an example, it is not a huge issue when a wrong suggestion is made by the algorithm on what music one can listen to [Shneiderman, 2020]. It is however more important when human lives or aspects of a human life are dependent on the decision that is being made. For example self-driving cars or pacemakers to this day still need a certain level of human control since they cannot (or should not) completely overtake the human role [Nof, 2009]. Therefore multiple different classifications have been made to distinguish the possible different levels of automation. Where table 4.1 shows the 10 levels of automation as stated by Sheridan and Verplank [1978] who were one of the first to define different levels of automation. Another classification has been made by Shneiderman [2020] who simplifies three different levels of application, based on the levels of automation by Sheridan and Verplank [1978]. The first, where the music genre suggestion example is part of, is named '*Recommender Systems*'. The second, which includes self-driving cars, is '*Life-critical Systems*'. Administrative tasks from governmental bodies fall under the third category, *'Consequential Applications'*, mixed with elements of the *'Life-critical Systems'* category, since these administrative tasks can have a huge impact on the lives of individual citizens.

Table 4.1: Levels of Automation by Sheridan and Verplank [1978]

Automation level	Description
1)	The computer offers no assistance: humans must take all decisions and actions.
2)	The computer offers a complete set of decisions/action alternatives, or
3)	Narrows the selection down to a few, or
4)	Suggests one alternative, and
5)	Executes that suggestion if the human approves, or
6)	Allows the human a restricted time to veto before automatic execution, or
7)	Executes automatically, then necessarily informs humans, and
8)	informs the human only if asked, or
9)	informs the human only if it, the computer, decides to
10)	The computer decides everything and acts autonomously, ignoring the human

Gulenko et al. [2020] mention another division of levels of automation ranging from a completely manual level up to full automation with only 3 levels in between called *managed*, *predictive* and *adaptive*. Although this distinction is mainly focused on AI implementation in IT systems, it again provides a different way of looking at the level of automation. And it shows that the type of task where the AI is used is again an important aspect when deciding upon the different levels of automation. Automated Decision-Making (ADM) tools, such as AI, can range from supporting decision-making to autonomously taking the decision which is considered as full automation of that decision-making process without human involvement [Araujo et al., 2020].

Focusing on administrative tasks in the public domain, Roehl [2022] has created two different tables of the level of automation. The way that decision-making processes are organised nowadays has changed immensely over the past few years and the rise of new supportive tools and technologies such as AI and other decision-support systems have played a significant role in this transition [de Boer and Raaphorst, 2021]. The level of full automation, where there is no more human involvement seems too far-fetched for now due to safety issues, social security, and responsibility but also due to the sole fact that it is not yet possible for most decision-making tasks [Gulenko et al., 2020; Shneiderman, 2020; de Boer and Raaphorst, 2021]. However, even within decision-making tasks, one can distinguish different types of tasks. Some may be more drastic for citizens and some may be less complex to complete. Thus, de Boer and Raaphorst [2021] conclude that low-level complex tasks can be more easily fully automated by a support tool or software such as an AI algorithm, because less human involvement is needed here. Therefore this research aims to look at a less complex task of civil servants and a more complex task to check for the differences in perceptions of the respondents.

Roehl [2022] compares multiple classifications considering the level to which a certain task can be automated where he simplifies these classifications to three levels: *"No automation, Semi-automated decision-support ("in-the-loop") and Fully automated decision-making ("out-of-the-loop")"*. For this research, the main focus lies in the middle level, the semi-automated decision support. It is here where different types of interaction can take place between humans and AI and where the AI becomes more autonomous and the human more of a controller and less of a decider. The lowest level where there is no AI involvement can be seen as the current situation (needless to say, that in some instances, pilots are now running on administrative tasks aided by AI). A situation where an AI is fully in control of public administrative tasks and the human is completely out of the loop is not expected to become a reality in the near future as already mentioned before. The community of municipalities (VNG) in the Netherlands has issued new principles on how we should deal with AI in the public domain and that it is for now inevitable to have a human in the loop [Kokkeler, 2022]. With the response from the ethical committee of the municipality of Eindhoven, it is clear that many of those principles still need further debate and that full automation of administrative tasks is not realistic in the short term [Kokkeler, 2022]. Therefore the focus lies on identifying criteria that could help in attaining a responsible hybrid human-AI system.

Six dimensions were classified by Roehl [2022], based on his comparison of classification, where authority shifts from the human or civil servant to the AI algorithm or the technology. These different dimensions are stated in table 4.2 below and are also subdivided into the 3 dimensions mentioned above.

Table 4.2: Levels of automation by Roehl [2022]

General Type	Specified Type	Explanation
No Automation	1) Minimal Automation	Almost fully entrusted to humans.
Semi-automated	2) acquisition 3) suggested 4) supported	Human asks the AI for suggestions. AI suggests appropriate further steps. AI suggests a narrow range of decisions
Fully automated	5) Automated decisions 6) Autonomous decisions	AI takes the decision and informs the human. AI takes the decision and does everything.

Based on the case study of using an algorithm such as that of BAIT by Councyl in the decision-making of civil servants, a slightly different classification than in table 4.2 is made here. This is done due to the limited amount of time and the levels of interest regarding BAIT. For every AI and task, some levels of automation could be added or deleted since these could be of more interest when looking at a particular AI and/or task. In this case, the 6 levels as stated by [Roehl, 2022] are slightly adjusted and decreased to 5 levels of automation, including a level of no automation, where no AI is present. These levels can be seen in table 4.3.

Table 4.3: Levels of automation specified for this case study

General Type	Specified Type	Explanation
No Automation	1) No Automation 2) Introspection	No AI involved Almost fully entrusted to human.
Semi-automated	3) Routing 4) Advisory	AI indicates the more complex ones, human decides. AI advises the human before taking a decision.
Fully automated	5) Decision Support	AI takes the decision and informs the human.

These different levels of automation will be used in the rest of this report and referred to according to the specified type. The results will indicate how policymakers make decisions based on certain criteria and requirements between these different levels and how these levels differ from each other in their impact.

The definition of these levels is defined here. In the case of no automation, this is assumed when no decision support tool is present in the decision-making process and everything is done manually by the civil servant. For the level of introspection, the AI is being used in the decision-making process but merely as an informative and learning tool to assess its own state McCarthy [2007]. It does not intervene with the decision-making at that moment though it could be used for assessment purposes and therefore indirectly have an impact on decisions that will be made in the future. An advisory level of automation is the first level where humans and AI both contribute to the decision. In this case, the human still takes the decision though it can see what the AI would have advised him to do. The human can learn from AI but still has full authority in the decision-making. The routing level of automation is the other way around. First, the human gets to see what the AI would advise before making its own decision. This results in AI having an influence and impact on the decision made by the human. When looking at the more automated levels. The decision support level is close to full automation though the only difference here is that the human is still involved without having to make a decision. The AI takes the decision though the human is still present and can thus intervene or notify when the person is in doubt of a decision for example.

Full automation has been kept out of this scope. In the case of full automation, the human is ”out-of-the-loop” and is no longer needed for the decision-making process. The AI takes the decision completely autonomously. This is however illegal in most cases and is not feasible nor is it wanted in the near future. Since AI is often a black box, meaning that one cannot retrieve why a certain outcome has been given by the AI, key elements of the protection of legitimate interest in the decisions making process can not be reached. Examples of these are the *'obligations to provide reasons'* and to possibility to check for *'proportionality'* and *'reasonableness'* [Carlizzi and Quattrone, 2023].

4.2 CRITERIA FOR THESE LEVELS OF AUTOMATION

This section focuses on the criteria that determine the level of automation. These constitute criteria that levels of automation can score differently on. Based on these different wanted outcomes one could determine a set of criteria that results in a level of automation. Though there are multiple criteria present, as mentioned in the chapter 2, the main criteria for this research will be taken further into the investigation.

A lot has been written about values and how to quantify those when looking at the digitalisation of decisions in both the public and private domains. Based on this literature, this section suggests a classification where those values could be considered as overarching concepts for the criteria that in their turn determine the level of automation as can be seen in table... For each criterion, the importance to the civil servant will also be briefly explained. As mentioned in the previous chapter this research will focus mainly on the stakeholder that is the user of this new technology and that is supposed to be supported by this new technology. As [Engin and Treleaven, 2019] mention how important the inclusion of civil servants is in order to have a safer and better functioning implementation of these automation levels. These different levels of automation all have a different intervention level by a human expert that should be well balanced for that particular decision Engin and Treleaven [2019]. This section will provide insights into the different criteria and how one should address those from a civil servant perspective in order to get to a well-balanced level of automation.

4.2.1 Efficiency Gain

Hood [1991] already mentioned in 1991 the increase in efficiency through automation was one of the main driving forces behind this new technological change. E-government tools or GovTech appliances where governmental tasks such as decision-making on public matters and interests can result in great efficiency benefits such as higher quality, higher productivity and faster service [Ranerup and Henriksen, 2019]. As Ranerup and Henriksen [2019] conclude, more research is needed on public decision-making and discretion related to automated decision-making. Looking at multiple case studies at different levels of local government and executing governmental organisations is vital in order to get a full and complete picture and understanding of the different applications and the different levels of automation for different tasks and different settings. The use of new technological innovation such as the use of AI in the decision-making process has an enormous impact on operational efficiency, individual innovation and employees' creativity [Satispi et al., 2023]. Different levels of automation can result in operational efficiency gains as it can overtake human jobs to a certain extent differing per kind of decision-making process as Coombs et al. [2020] conclude. They also mention the need for further research in a multi-disciplinary approach through different methods to get a complete picture as possible of this complex problem (e.g. through a qualitative case study, stakeholder survey, observations, focus groups, etc.) Coombs et al. [2020].

Looking at the different levels of automation, the efficiency gain can be different for each level of automation as one can imagine. Hood [1991] already mentions the increase in efficiency gains the higher the level of automation and the less human interaction there is. A machine can make a decision based on a few criteria a lot faster than a human. The question is, can it take into account the context, the meaningful aspect, and the human side? Especially when a decision is being made on a very sensitive topic such as whether to financially aid a citizen in need or in the case of child placement. As a recent study concluded from analysing the potential of the new and highly-advanced AI tool, Chat GPT, in dealing with safety-related issues that AI is not able to understand the context in certain situations, as in their research it made incorrect or potentially harmful statements [Oviedo-Trespalacios et al., 2023]. Ethical considerations and safeguards need to be ensured in order to implement safe, responsible automation of human tasks.

When looking at the implementation of new technologies that can take over human tasks partially or completely, efficiency is the main reason and driving force of this automation, also in the public domain [Mehr, 2017; Muhlenbach, 2020; Hood, 1991; Nau, 2009]. Looking at levels of automation, the level of efficiency also decreases or increases depending on the level of automation that has been selected as a strategy. A lot has been written on how well an automation tool such as ML, NN or AI, can overtake a task of a human and how accurate this decision support tool will be [Nau, 2009; Bannister and Connolly, 2020; Martinho et al., 2021].

Hood [1991] mentions already the importance of efficiency gains and ethical values in automating public services mainly in information technologies. Though safety should be safeguarded. There is a trade-off here that has aspects of safety, responsibility, explicability, transparency, and fairness [Lindgren et al., 2019; Tangi et al., 2022]. Which all reflect the role of the human in the decision-making process. The human working with these algorithms needs to be able to explain the decision, show how one came to this decision, bear responsibility if a wrong decision has been made, and make sure that the decision results in a safe outcome, that the outcome is safe.

Thus the role of the human and the tasks that the human has and the requirements the person has to fulfil can all be placed below the criteria of human control. Human control is a huge topic of interest these days especially related to AI. Plenty has been written on the concept of 'meaningful' human control, especially in the domain of autonomous weapons. This criterion will be discussed in the next section.

4.2.2 Human Control

As can be resulted from the previous section, human control and efficiency gain are highly intertwined since a new technology such as AI may result in more creative innovations from the civil servants though less human control over a decision may be in place. Another inescapable criterion in choosing a strategy for the level of automation of human tasks by decision support tools is the level of autonomy or human control. This differs per strategy and per level of automation and is dependent on many factors such as the type of task, the kind of data, and the technical and human possibilities. Its main trade-off can be found with efficiency, e.g. when a task is more digitised it is faster and thus could be considered more efficient, as became clear from GovTech appliances [Engin and Treleaven, 2019; Bharosa, 2022]. Since we would often like to have a human-in-the-loop, especially in highly sensitive cases, this might limit the increase of efficiency. A now famous example of this trade-off and the coming risk with automating sensitive tasks is the so-called "Toeslagenaffaire" in the Netherlands. Here, sensitive data and high-impact decisions were made through algorithms that were used to check for frauds [Hadwick and Lan, 2021]. Resulting in wrongful fraud detection, a decision that was inhumane and not respecting the rights of an individual.

As Siebert et al. [2022] also mention the huge importance and impact of meaningful human control where responsibility is a key aspect. They created four properties that are important for ensuring meaningful human control and are a beginning of operationalising these criteria further [Siebert et al., 2022]. This research aims to dive into this criteria even further and to determine what factors impact the trade-off between human control and efficiency. Apart from the four criteria mentioned by Siebert et al. [2022], what other aspects play a role in determining the wanted level of automation and how does this relay to efficiency gains or losses? Not to forget that the level of human control and the way that this is implemented is also a huge issue of security. This has often been mentioned as being an important factor also looking at the trade-off with efficiency in human-algorithm collaboration [Terra et al., 2020]. So one needs to keep this aspect in the back of their mind when looking at the trade-off between efficiency and human control.

Different conditions have been discussed in the literature to reach meaningful human control Siebert et al. [2022]; de Sio and van den Hoven [2018]. In order to reach meaningful human control, there must be adhered to values such as responsibility, accountability and transparency [**?**]. When determining the level of automation and gaining insights into the trade-offs between the different criteria, it is important to address these values on different levels as well. As was also shown in figure 2.1 by Wirtz et al. [2020] on the challenges of AI. Methnani et al. [2021] suggests a dynamically adjustable level of automation (autonomy), where these values need to be addressed in the changing context and regarding the (required) knowledge and ability of the human. This dynamically adjustable level of automation is interesting since one can imagine that, when situations change over time, a certain task may be preferred to be less automated than previously accepted. Responsibility, autonomy and transparency are three key values that are related to the criteria (meaningful) human control that will be further discussed and researched in this case study research. Autonomy is also present in the discretion of decision-making, which cannot be delegated to an AI, since this may result in a lack of moderation of power Carlizzi and Quattrone [2023]. As de Boer and Raaphorst [2021] mention, street-level bureaucrats are the only ones with autonomy in traditional bureaucracy, where the rise of new IT systems alter this decision-making process and the implementation of public policy.

4.2.3 Quality

Quality aims are often linked to being more accurate and more constant in less time and supporting the human as much as possible in order to have the human do its job "better" [Kuziemski and Misuraca, 2020b]. AI can help in decision-making becoming more concise and consistent thus adding to the quality of the decision-making [Kuziemski and Misuraca, 2020b; Ivanov, 2022; Andronie et al., 2021]. The higher the level of automation the higher the quality. But what if context factors are not taken into account and the decision is fully automated? Every decision may be taken according to the same criteria with no deviation in the interpretation and opinion of the civil servants, but the context disappears where an AI cannot give meaning to a certain scenario or an exception. This may result in wrong decisions, thus one could say that quality increases with the level of automation but to a certain extent. Quality is a hard-to-define concept as Marjanovic and Cecez-Kecmanovic [2017] shows. As is the case in automation where subjectivity also plays a role. What is good or good enough, or consistent enough? Coombs et al. [2020] explains that organisations focusing on human capabilities when implementing AI and not focusing on fully automated tasks, prioritise the quality of their service. These organisations have a higher probability to formulate an effective strategy. The level to which organisations should augment intelligent automation to increase the quality of their service is uncertain and unclear, though full automation is not likely nor wanted in the short term [Coombs et al., 2020].

4.2.4 Implementation Effort

The criteria of implementation effort relate more to the technological aspects of different levels of automation. Full automation means that the AI should also retrieve all the needed data, whereas a hybrid version could suggest that only the software of the AI needs to be compatible with the software of the municipality. In every level of automation, privacy and security should be huge aspects to take into account, but these are also part of the implementation effort for this research. Traceability of the decision and of the data used in a decision gets harder when the decision has a higher level of automation for example. This also has an impact on the deskilling of employees, when the implementation is not done thoughtfully. Coombs et al. [2020] also mentions that different levels of human-AI interactions emerge with different impacts on the way an organisation works and is organised. Shaw et al. [2019] elaborates on the challenges of *"meaningful decision support"* and *"explainability"*, where the implementation of AI needs to be meaningful in the sense that it should contribute to the civil servant and the citizen. When a civil servant can make it more clearly insightful in a quicker way this is for both parties beneficial[Shaw et al., 2019]. Though how the AI has been used in taking the decision needs to be clearly mapped in order to be able to explain the train of thought and how they (human-AI interaction) came to the decision [Shaw et al., 2019].

4.2.5 Deskilling

An often debated criterion is the deskilling of employees due to the automation of their tasks [Sambasivan and Veeraraghavan, 2022]. Because their tasks are being automated they do not have to be as qualified as they were before in order to fulfil their tasks. This however does also result in new knowledge required on the use of AI and working with an AI. Dhungel et al. [2021] mention the importance and possibilities of training and re-educating employees in working with AI, specifically in the civil servant's case where the outcome of the decision has an impact on (individual) citizens. An understanding of why, what or how the work of civil servants is impacted by the implementation of AI in whichever level of automation is of key importance to the good implementation of AI [Dhungel et al., 2021]. Rafner et al. [2021] mention the need for more empirical domain-specific case study work to dive into the opportunities for hybrid intelligence between human and AI, where deskilling can be transformed to *"reskilling"* and *"upskilling"*. Another example where it may be easier to understand the need for a hybrid version where deskilling does not occur is aviation, where in full automation pilots lose the operational skills to manually fly the aeroplane [Crespo, 2019]. A lack of situation awareness and diminished capabilities of the pilot can of course have highly unwanted outcomes [Crespo, 2019]. Thinking about the relationship between AI and humans in decision-making is key in automating decision-making processes [Sambasivan and Veeraraghavan, 2022; Rafner et al., 2021]. This research aims to provide this need.

4.2.6 Vulnerability

Plenty of risks and challenges come with AI such as the problem of bias in data but also the bias of those creating the AI algorithm, and the issue of privacy [Feuerriegel et al., 2020]. This section however tries to look at vulnerability as a little bit more practical. What are the practical vulnerabilities of implementing AI in a certain level of automation? As mentioned in the previous section, vulnerability is slightly linked to deskilling, where the vulnerability of an automation level, with AI more in control, increases. A lot can go wrong when the AI fails or errors occur. Going back to the aviation example, if in full automation the AI fails there will be more damage than in a hybrid level of automation where the human is not completely out of the loop [Crespo, 2019]. Vulnerability is also linked to the privacy of citizens in the case of applications in the public domain, for example on the way this data is extracted and how it is used and analysed [Lockey et al., 2021].

Another important vulnerability Lockey et al. [2021] mention is the paradox between automation and augmentation. Augmentation focuses on working with the machine while automation is defined as machines overtaking human tasks. Lockey et al. [2021] elaborate on the risk of working with an AI but letting it also overtake some tasks whereas this may need adjustment over time due to changing parameters and context. This may result in a completely new and different role for the civil servant in the case of AI application in the public domain where they not only work with the AI but also make sure that it is constantly in line with the current societal conditions and frameworks. Another aspect of vulnerability when using AI may seem more simple. When fully automating certain tasks by a machine or algorithm there is a risk of power failure where no decision will be made in the case of local governments. This also relates back to deskilling where the damage will be larger if civil servants will have completely relied on the AI and cannot make these decisions on their own anymore. The type of AI plays an important role here. As soon as all the steps that an AI follows to get to a decision are well documented, it would be easier for a civil servant to still follow what the AI is doing and understand how it comes to a decision. However, by automating decision-making, we are losing this process towards the outcome of a decision. This is starting to become more and more a theme of public debate, thus transparency and explainability again play an enormous role in implementing AI successfully [Februari, 2023]. This will also be beneficial to the acceptability by all stakeholders as will be explained in the next section.

4.2.7 Acceptability

Acceptability is also intertwined with many of the above-mentioned criteria and can be viewed differently from different stakeholder perspectives. The acceptance of citizens for local governments using AI may be different from that of civil servants as the operators of an AI or the policy-makers deciding on a level of automation. For example, when looking at the aviation case again, the policy-makers may find it very attractive and acceptable to introduce fully autonomous flights whereas citizens (clients) do not feel comfortable at all in flying with an autonomous aircraft [Crespo, 2019]. In the case of the human operator, the civil servant, it may be less acceptable to use an AI as a support tool in decision-making when the civil servant remains responsible for the decision formed through human-AI cooperation. A lower level of automation may thus be preferred and accepted by a civil servant. This research aims to give insights into the trade-offs that civil servants make that result in their acceptance or their willingness to use an AI in aiding them to make decisions more efficiently [Araujo et al., 2020; Delfos et al., 2022]. In decision-making by the government over citizens, it has already become clear that citizens do not accept AI algorithms that aid in decisions with ”real-life consequences for humans” [Kuziemski and Misuraca, 2020a]. As Siau and Kam [2006] mentioned, without full trust from the citizens, data collection and data use for AI automation won’t be as successful. Taking humans out of the loop for important decisions on the lives of individuals is unacceptable as stated by Kuziemski and Misuraca [2020a]. However, this can change over time, per the type of AI that is used and whether the civil servant knows how to use the AI where they are still responsible for the outcome of a decision and able to explain this to a citizen.

4.3 IDENTIFIED TRADE-OFFS

These criteria for the level of automation as explained above can have trade-offs with one another. When choosing a higher level of automation, with more AI influence in the decision, one criterion may score higher whereas another may decrease. Based on the wanted and desired outcome a certain level of automation can be picked. It is however insightful to get a clear overview of these trade-offs for different scenarios. Especially when looking at the civil servant, the user or operator of the AI, little research has been conducted on the trade-offs from their perception and on how domain experts view this. As human control has been mentioned often as an important criterion when determining a level of automation

As could be concluded from the desk research, one of the main driving forces for automated decision-making in the public domain is efficiency gain though the level of human control is also very important, most definitely from a civil servant’s perspective. The expert interviews aim to validate these main criteria where the perspective of the civil servant is investigated in the vignette experiment. Though this research considers criteria such as transparency and explainability to be more related to the type of AI and should be taken into account when deciding upon this topic. The criteria and the trade-offs that are discussed here are related to the choice of a certain level of automation with a particular AI. Taking into account the type of task and the context, knowing that this can differ over time. To what extent is efficiency gain wished for and when is it better to have more (meaningful) human control, in order to better address a situation and better understand the context and decide with a more human-centred approach? The level of human control is decreasing with the increase of efficiency when automating tasks, so a trade-off is present here and this research aims to determine a responsible level of automation based on these kinds of trade-offs. Though determining where, how and when to implement automation in a decision-making process is also very sensitive, the level of human control may still differ per level of automation considering where, when and how it is implemented. This will be explained in the sections below, elaborating on why human control and especially meaningful human control is such an important criterion according to the existing literature, frameworks and theories.

4.4 FRAMEWORKS FOR AUTOMATION

In the literature, two frameworks have been created in order to come to a level of automation as proposed by Sheridan and Verplank [1978]. These frameworks will be discussed below and aim to explain and substantiate the complexity and versatility of this problem. They will be used in order to provide guidance for those willing to experiment with the use of AI in automating decision-making by civil servants in the public domain. These frameworks focus on particular aspects of the decision-making process that are considered to be important context factors in this research. Especially in automating different tasks or decision-making processes, context factors are very important to take into account as certain values and criteria can drastically change based on the context of the decision or task [Parasuraman, 2000; Wickens and Dixon, 2007]. Understanding the context is very important as these socio-technical innovations as automation are embedded in the context of decision-making and interact with it. Some of these context factors will be used in the interviews and the vignette experiment in order to gain knowledge of the trade-offs between criteria and context factors and how these impact each other. They are also highly needed to provide a realistic case and scenario in the vignette experiment.

4.4.1 Flow Chart for Level of Automation

Parasuraman et al. [2000] created a model building on the ten levels of automation as created by Sheridan and Verplank [1978]. They proposed four different classes where automation can be implemented in order to gain efficiency or reduce costs. Related to decision-making these four different classes could be defined as the following.

1. *information acquisition*; where automation can be used in gaining the data in order to make a decision
2. *information analysis*; where automation can be used in analysing the data that has entered the decision-making process, resulting in possible advice, summaries, emergent features, etc.
3. *decision and action selection*; where automation can be used in the actual decision-making based on the advice or summary from the previous step of information analysis.
4. *action implementation*; where automation is actually used in the execution of the decided action.

As the scope of this research aims to provide guidance in automating the decision-making process and not the data gathering or the execution of the decision the *information analysis* and the *decision and action selection* are the main important. Parasuraman et al. [2000] provided a flow chart which helps in understanding what the automation should be used for, what part of the decision should be automated and focuses on primary and secondary criteria that determine how well a level of automation would fit for a specific decision-making task, as can be seen in figure 4.1. These two levels of criteria will be important for the rest of this research as these could be considered to be performance criteria, but also context factors. The next section focuses on the model by Cummings and Bruni [2009] who built upon the framework by Parasuraman et al. [2000].

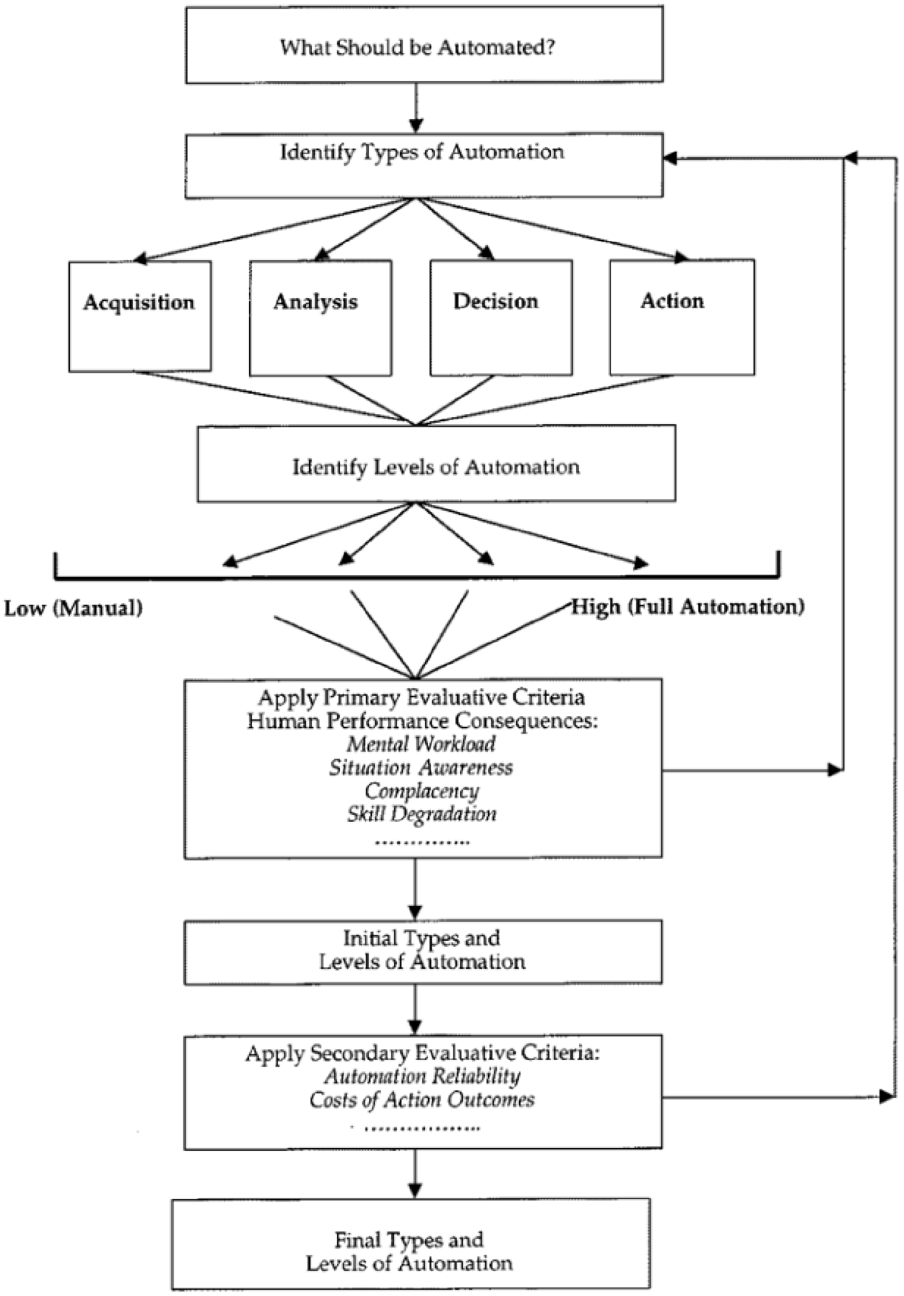


Figure 4.1: Flow chart showing the application of the model of types and levels of automation, adapted from Parasuraman et al. [2000]

4.4.2 Human-Automation Collaboration Taxonomy

Cummings and Bruni [2009] created a framework called Human-Automation Collaboration Taxonomy (HACT), also extending the different levels of automation as stated by founding fathers Sheridan and Verplank [1978], ranging from fully manual systems to fully automated systems. Looking at the 10 levels of automation in table 4.2, the first four are the more human-centred levels where execution is not part of the algorithm but still belongs to the tasks of the human. Based on the papers by Parasuraman et al. [2000] and Cummings and Bruni [2009] established a framework where three "collaborative decision-making process roles" are present. Mainly focusing on the automation of the information analysis and the decision selection processes [Cummings and Bruni, 2009; Parasuraman et al., 2000]. These three roles include the *moderator, generator* and *decider* and can each have a certain level of automation that has a different impact on the decision-making process and the end result. These three roles can be automated in 5 levels ranging from the human who takes the final decision to a system where the algorithm or AI takes the decision. This shows the importance of the different roles that one human can have in the same decision-making process and where automation can support and aid the human in decision-making. This is important for this research as it is important to be aware of the different human roles that could be automated with different levels of automation in the decision-making process by civil servants in local governmental organisations. The three different roles resulting from the framework will be explained below.

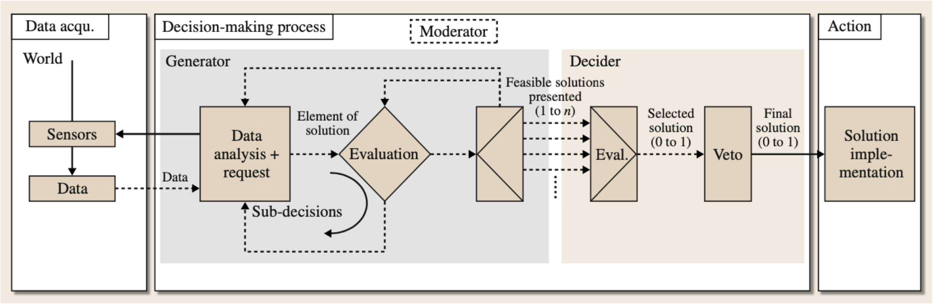


Figure 4.2: The three collaborative decision-making process roles: moderator, generator, and decider, adapted from Cummings and Bruni [2009]

Moderator

The moderator takes care of the entire process and makes sure that no step is missed or skipped, it makes sure that the decision process doesn't halt and takes the context factors into account. This is for example time pressure, but can also relate to context factors in the current situation where a decision has to be made. For example, in the case mentioned by Cummings and Bruni [2009] on the automated antimissile system, in a scenario where there is no war, it will be very unlikely that a missile would enter your airspace but that it is more likely that a friendly aeroplane has entered it. Understanding context is therefore very important and the human perspective, although possibly biased, may be crucial here.

Generator

The generator is busy analysing and evaluating the requirements and coming up with possible solutions [Cummings and Bruni, 2009]. This entails taking into account what requirements should be adhered to and what laws and regulations should be regarded. The possible solutions are possible outcomes when taking this decision. This task could be automated when an AI knows how to trade off these requirements and how to value them.

Decider

The decider decides upon the decision based on the solution the generator proposes based on their evaluation of the requirements and criteria [Cummings and Bruni, 2009]. The most important aspect of this role is the veto power, meaning that as soon as the decision has been taken by the decider, no one else can change or supersede this. All three of those tasks can be done by a civil servant, by an AI or by a combination of the two where humans and algorithms collaborate [Cummings and Bruni, 2009].

HACT for Civil Servants

Looking at the case of civil servants at local governments taking decisions that impact individual citizens, these three different roles are interesting to investigate. Especially when trying to explain the trade-offs that occur when determining a level of automation from a civil servant's perspective and thus also the purpose of a level of automation has an impact on this trade-off. This will most likely differ per type of decision where different context factors play a role and where these three roles will fluctuate. HACT makes use of these levels of automation for these different roles and combines different levels of automation in order to provide the wanted service and/or outcome.

Cummings and Bruni [2009] mention, however, the importance of the social impact that human-automation collaborations have. This is especially the case in the public domain where decisions by civil servants impact the individual lives of citizens. Quite some experimenting has been done with pilots on the implementation of AI in the public domain focusing on the information acquisition side and in smart city appliances [Parasuraman et al., 2000; Fiebag, 2022; Gornishka and Sukel, 2022]. The implementation of AI in the decision-making and analysing of data will therefore be the main focus of this research and will be done through interviews which will be analysed in the next chapters, in order to come to recommendations that could be used in determining the responsible level of automation.

4.4.3 Where to Use an Algorithm

The previous two sections show existing frameworks on where to use algorithms, and in what stage of the decision-making process. Where you decide to implement AI is also dependent on certain criteria and for each type of automation that is suitable (acquisition, analysis, decision, and action [Parasuraman et al., 2000]) different levels of automation could be determined that would provide a responsible implementation of AI. It is therefore important to be aware of these different ways of implementing AI. Before implementing AI, think about where the possibilities lie to implement AI, what the benefits, downfalls and challenges are when implementing a level of automation in this part of the decision-making process. But also think about where a responsible level of human control should be present. Should it be at the action phase of the decision, or preferably at the decider or the generator in the decision-making process? What is the impact of these different levels of automation on the amount of human control that is still present? Who is finally responsible for the decision made by the partially automated decision-making process? Resulting in meaningful human control, with a responsible implementation of AI with clear roles and responsibility assignments.

Cummings and Bruni [2009] mention the impact of certain context factors on the preferred presence of a human being in the loop. Context factors such as the sensitivity of the decision, impact on the individual, amount of risk and the time, within which the decision has to be made, are of high importance to the wanted level of human control and human autonomy that is present in a certain decision-making task. A higher level of human control may be preferred when choosing a level of automation in the decider role in decision-making than when automating a moderator role because the context in which the decision is made, is more important than the decision-making process where an algorithm can easily moderate if everything is being done according to protocols. These three roles will not be extensively discussed in the rest of this research though, for the bigger picture of implementing a responsible level of automation, it is important to be aware of all the different ways, forms and shapes in which automation in decision-making can take place. By integrating these frameworks into the decision-making landscape, organizations can navigate the complexities and challenges of automation, harnessing the benefits of technology while upholding ethical standards and ensuring public trust. The following section touches upon the safety and control aspect that could also benefit to this public trust.

4.5 THEORIES ON CIVIL SERVANT CONTROL

This section dives into a theory that can be used in order to define the concept of safety and control and what context factors play a role here. Therefore, especially focusing on the role of the civil servant and how human control should be taken into account. This theory has been used by Dobbe [2022] on the implementation of AI in the public domain and to determine how we should deal with risks and safety also in terms of responsibility and impact on the decision. It discusses aspects of safety and how to reach 'Responsible AI', focusing on the role of the (human) controller.

4.5.1 System Safety Theory

Dobbe [2022] uses the System Safety Theory (SST) to draw lessons that could be applied to the implementation of AI. These lessons will be used to discuss the responsible use and implementation of AI, especially focusing on the aspect of human control, where these lessons should be tested with public servants in the public domain. One of the most relevant lessons to this research is the focus on the organisational culture within the public domain, where the organisation and the institution should be responsible for a saver culture, where human control is focused on implementing algorithms in decision-making where it can support the civil servant safely and come to a fair decision. The SST has been mentioned and extensively discussed by Laveson [2012] who mentions the need for a safety-focused perspective on the implementation of new technology where Dobbe [2022] uses this for AI implementation. A main aspect of the SST is the level of control that should be present when safely implementing engineering tools. According to the definition of control in open systems by Laveson [2012], the decision-making process by civil servants in the public domain requires four conditions in order for the controller to successfully and safely control the process. This is also shown in a standard control loop in figure 4.3 adapted from the book by Laveson [2012]. These four conditions are the following and could be applied to the criteria of human control looking at when human control is meaningful and to what extent this should be present in automating a certain decision-making process or a part of this process. These conditions are stated below and adjusted to the civil servant who is the controller of the decision-making process in one way or another. This concept of SST implementation was also mentioned in one of the expert interviews (Interview B.2).

1. *Goal Condition*: The civil servant must have a goal.
2. *Action Condition*: The civil servant must be able to effectively affect the system (Actuators)
3. *Model Condition*: The civil servant must have a model of the behaviour of the decision-making system.

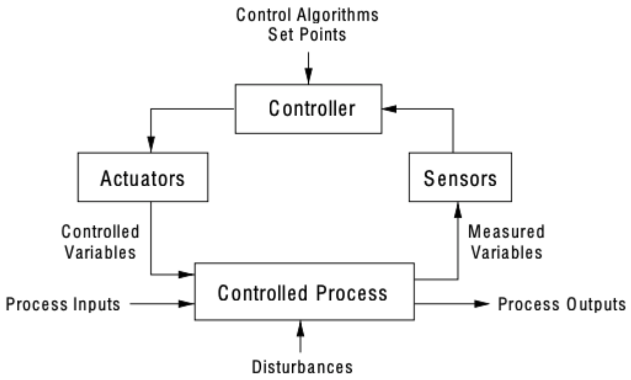


Figure 4.3: A standard control loop, adapted from Laveson [2012]

4. *Observability Condition*: The civil servant must be able to ascertain the state of the system, through feedback, measurements and observations, as Dobbe [2022] mentions (Sensors).

Including this notion of the SST by Laveson [2012] it is necessary to see that the role of a controller is incredibly important to ensure the safe implementation of new engineering tools. This can for example be applied to the implementation of AI in the public domain, as analysed by Dobbe [2022]. Dobbe [2022] concludes that this theory cannot solve us in helping safely implement AI straight away, but it can help us in determining *when* and *how* to implement AI. And what the role of the civil servant should be in terms of human control when automating decision-making by civil servants in the public domain. Therefore this research focuses on how to determine a responsible level of automation, especially when looking at the criteria of human control and when this is considered to be meaningful and responsible. Aiming to provide guidance on a more safe, accepted and responsible implementation of AI in the public domain.

4.6 CONCLUSION

This chapter discusses the different theories and frameworks that have been created on automating decision-making and on the criteria that come into play when deciding upon a tool for automation (AI, ML, IoT, NN, etc.). Multiple classifications have been made, each providing valuable insights into these different levels of automation that could be used for digitalising governmental institutions, which is one of the main challenges that we face nowadays. Making sure that citizens all get equal, fair, safe and good services provided by the government, and staying ahead of the dangers of developing new technologies, is key. The models described above will be the foundation for this research, though they still require more testing in the public domain. The focus of these models has been on automating the decision-making and tasks of humans in all kinds of domains, such as air traffic control systems, GPS, and anti-missile systems. However, the implementation of AI in the public domain where decisions are being made over the future of civil servants has a different character.

This chapter explored two frameworks for automation in the context of decision-making by civil servants in the public domain. The Flow Chart for Level of Automation by Parasuraman et al. [2000] identified four classes of automation implementation, with a focus on information analysis and decision and action selection. Cummings and Bruni [2009] Human-Automation Collaboration Taxonomy (HACT) introduced three roles - moderator, generator, and decider - that can be partially or fully automated. Understanding the different levels of automation within these roles is crucial for responsible implementation.

Considering the implementation of AI, it is important to determine where algorithms can be effectively used in the decision-making process. Factors such as the sensitivity of the decision, individual impact, risk, and time constraints play a significant role in deciding the appropriate level of human control. The decision on where to implement automation should consider the benefits, challenges, and responsible human involvement. The presence of human control at different stages, such as the action phase, the decider role, or the generator role, will impact the overall level of automation and the degree of meaningful human control. It is crucial to establish clear roles and responsibilities to achieve meaningful human control in decision-making processes where AI is partially automated. To get to a responsible level of automation, the trade-offs between the criteria need to be made insightful.n chapter 5.

In conclusion, this chapter provided insights into the frameworks for automation, the roles of human-AI collaboration, and the importance of meaningful human control and the SST in responsible AI implementation. The findings inform the subsequent chapters, which will delve deeper into the trade-offs and criteria associated with automating decision-making by civil servants in the public domain through interviews and a vignette experiment.

5 RESULTS

This chapter explains the interviews that shape the case study in the vignette experiment that is being conducted on two specific decisions that civil servants working for municipalities in the Netherlands make regarding special welfare for citizens. Based on the desk research on the existing classifications of levels of automation in the previous chapter, input was gathered for the interviews on the criteria and demarcations on context factors have been made for the vignette experiment. First, the set-up of the interviews will be discussed together with the responses of the experts. Next, the vignette experiment will be discussed, touching upon the goal of this survey experiment before explaining the different cases and scenarios that have been tested. The results of the interviews and vignette experiment are also presented here, aiming to provide insights into determining a responsible level of automation based on the made trade-offs by the respondents and their reasoning.

5.1 EXPERT INTERVIEWS

Based on the previous chapters, three research topics of interest could be determined that have been chosen to focus on regarding the rest of this research and mainly in these interviews. These topics include the criteria that experts consider when asked about determining a level of automation where the human and AI will take the decision together in a hybrid human-AI interaction. The second research topic is focused on the contextual factors that impact the level of automation or that are impacted by the decision for a certain level of automation. The third research topic is overlapping slightly with the previous one but focuses more on the responsible side of AI implementation and what the role of the civil servant should be when implementing AI at any level in the decision-making process. This last part also asks experts about the lacking knowledge in the current situation of the policy-makers when considering the implementation of decision-support tools such as AI. The data from these interviews have been coded according to a number of categories. Each of these categories will be discussed here using quotes and references from the interviewees. These different categories will also be presented with their coded subcategories in clear tables in this section.

Every question was stated in the context of automating the decision-making of civil servants of decision that directly impact individual citizens, for example in decision-making by civil servants in municipalities. The concept level of automation was clear for most interviewees though they were provided with a definition where full automation will be left out of the scope of this research as this is considered irresponsible for now and is hard to implement following the current laws and regulations, based on input from interviews and desk research. The interviewees were also informed that automation could be a tricky term to use since this research focuses on hybrid decision-making where AI and humans interact and together come to a decision. The role of the civil servant and AI should be determined in order to take a step towards a responsible level of automation. In these interviews, as well as in the vignette survey, the terms AI and algorithm are used intertwined. Algorithmic management has been a topic of research for the European Commission for the past few years under which AI is also stored as an algorithm [EC, 2019; Algoritmeregister, 2020]. Both terms are used to more clearly visualise what certain AI algorithms can do, and not to speak of AI only as Hal 9000 ('2001: A Space Odyssey') as mentioned in the introduction.

5.1.1 Reasons for Automation

From the literature, the main reason for using automation tools such as AI is considered a gain in efficiency [Hood, 1991; Mikalef et al., 2022]. Though this has mainly been seen with implementation in the private domain, evidence exists of this being the same driving force for automation in the public domain [Parasuraman et al., 2000]. The interviewees were asked what they considered to be the main reasons for public organisations to automate decision-making in any hybrid form where more autonomy moves from the civil servant to the AI for each higher level of automation. Their responses could be summarised into three main reasons of which increase in quality and efficiency gain were considered to be the main driving forces, followed by some reasons that had not been found in the desk study. These three main reasons for automated decision-making in the public domain were mentioned by most of the experts and could slightly overlap:

1. *Efficiency*: Faster decision-making (more decisions per time unit)
2. *Quality*: More consistency in the decisions made
3. *Quality*: Less bias as the decisions are made with less personal values from the civil servant

The latter two could both be categorised under the category of *Quality*. Though with most of these reasons often a 'but' followed, meaning that a lot of requirements needed to be met on the way this automation will be implemented in order to reach these increases in either quality or efficiency. As with almost every complex problem, it depends on many other aspects. One interviewee from the private domain mentioned the need for a good implementation if you want to reduce bias and subjectivity of the decision maker (Interview B.6). But the question quickly rises of what is *"good implementation"*, where the rest of the questions will try to give an answer to or at least some guidelines. Another interesting category that resulted from analysing the reasons for automating decision-making is the opportunities that it offers where one interviewee from the scientific domain highlighted the benefit of recognising patterns in the decision behaviour of civil servants (Interview B.8). This interviewee stated the importance of civil servants *"creating awareness of their own patterns and making them reflect more critically on their own decisions and thoughts"* (Interview B.8). In essence, using AI to provide *"a mirror"* to the civil servants could eventually increase the quality of the decision and better support the civil servants, *"decreasing their arbitrariness"* (Interviews B.3 & B.5). *"Showing the blind spots of the civil servants and highlighting the cases that need extra intention"*, which is stated as *routing* in table 5.8 and can improve the decision making by civil servants (Interviews B.8 & B.9). Other often mentioned main reasons for automating are the standardisation of certain tasks which could benefit both the quality of the decision (in terms of consistency) and the efficiency of the decision-making process (Interviews B.2, B.3 & B.4). Financial benefits were mentioned by two of the three private domain experts as a more general incentive for automating through the implementation of AI but not necessarily in the public domain. The coded answers for reasons to automate are presented in table 5.1 below.

Consistency is mentioned by a few interviewees as a reason for automating decision-making as this could be beneficial for the quality of the decision, where this is often wanted related to the goal of equal service and equality to all citizens (Interview B.7 & B.8). For this to be the case, one needs a clear definition of equality and a mutual understanding needs to be present on how to deal with these definitions, the same applies to "good" implementation. Once we determine what is meant by

Table 5.1: Reasons for Automating Decision-Making, Coded Responses

Category	Public	Private	Scientific	Total
Efficiency Gain	3	2	2	7
Consistency	2	2	3	7
Less Bias	2	1	3	6
Civil Servant Support	2	1	3	6
Arbitrariness	2	1	1	4
Complementary	1	0	1	2
Interpretability	0	1	1	2
Corruption	1	0	0	1

these kinds of values, the policymaker can dive further into the applications of AI at a responsible level of automation. A more practical, reason that was mentioned for automating decision-making by civil servants in the public domain is the rise of big data and the increasing amount of (recurrent) decisions that have to be made by these civil servants. Often tasks are more prone to being automated when *"a lot of data is present* (Interview B.4) *"of millions of people"* (Interview B.2) and where it can consequently *"save a lot of human work* (Interview B.3). And as was mentioned in Interview B.9, economies of scale is an important practical aspect here as well, as we do not have *"a sufficient amount of civil servants"*. AI can be a solution for these ever-growing amounts of data and a limited number of personnel by quickly analysing these cases and providing a sufficient answer (Interview B.6).

A final reason for using AI in decision-making by civil servants is the complementary aspect of civil servants and AI. It is important here to make sure that both AI and civil servants work together and help each other. "Best of both worlds" (Interview B.9). This was already mentioned when highlighting the blind spots of the civil servant. But there should be a balance in the hybrid cooperation between humans and AI, and this may again be dependent on other factors such as contextual factors, which will be discussed later in this chapter. When "computer says no", it does not mean the civil servant should also say no (Interview B.9). The civil servant should be stimulated to think about what decisions they make, what the impact is, why they make that decision, what patterns occur in their decision-making and how to deal with this. This could also help in avoiding personal bias and arbitrariness from the civil servant (Interview B.6. When implemented correctly this may result in higher quality decision-making and still with a high level of human touch and understanding of context. Where civil servants complement, in their turn, the AI. After being asked for the reasons for choosing automation, the interviewees were asked to think of criteria that could be given a score for different levels of automation and how these could possibly trade-off. These criteria will be discussed next.

5.1.2 Criteria for the Level of Automation

The interviewees were asked what criteria they would find important to take into account when deciding on a certain level of automation. These criteria and how the policymaker would view those criteria would determine or at least point in the direction of a certain level of automation. From the criteria that resulted from the desk research, a few were also mentioned repetitively in the interviews. The criterion of human control will be discussed separately as this resulted as the main criterion from the literature and was also mentioned by almost all interviewees. All the other mentioned criteria for the level of automation are stated below in table 5.4.

Table 5.2: Criteria for level of automation, Coded Responses

Category	Public	Private	Scientific	Total
Efficiency Gain	3	2	2	7
Impact	2	2	2	6
Quality	2	1	1	4
Vulnerability	1	2	1	4
Context-Dependent	1	1	2	4
Acceptability	1	0	0	1
Deskilling	0	0	1	1
Corruption	1	0	0	1

A number of criteria, according to the interviewees were not yet present in the list that was discussed in chapter 4. But before these are discussed, first the criteria from the desk research are validated with the information from the interviews. Efficiency is mentioned by many of the interviewees as a driving force of automation but also as a criterion, as with higher levels of automation, they envision a larger efficiency gain. Here it is often mentioned in the context of easy-to-automate decision-making tasks, for example when there is clear guidance from laws and regulations (Interview B.2, Interview B.4, B.7 & B.8). Quality is also mentioned as a criterion for choosing the level of automation. If the quality does not increase, the interviewees do not see the value of automating the decision tasks, unless it increases the efficiency a lot ((Interview B.4). Another criterion that is often mentioned is the vulnerability of the citizens and whether automating this decision does not impact the citizen too much in a negative way (Interview B.4). The efficiency gain is also mentioned in Interview B.8 as looking at the repetitiveness of the decision, however, this is not seen as a criterion for the level of automation but is more considered as a warning. *"For those decisions that occur on a repetitive notion, it is important to ask the question why do they occur so often?"*. The interviewee highlights here the importance of looking at the reasons why a decision has to be taken more often. Is there another problem with why this decision is taken so often? The reflection for the civil servant is super important and if automation can benefit by providing "a mirror" to the civil servant, then this could be beneficial for the efficiency but mainly for the quality of the decision-making process and of public service in general. Where quality goes further than accuracy or loss, as is often understood in the scientific literature (Interview B.8). This also includes context dependency, although this was mentioned in the interviews as a criterion, this research takes this category apart to analyse contextual factors in the next section.

Acceptability was mainly mentioned as a criterion for choosing a responsible level of automation when regarding the acceptability of citizens. This was considered very important together with the impact on the citizens, which will be discussed in the next two paragraphs. Efficiency gain could also be measured as a decline in waiting time for the civil servant which will build on the acceptability of the citizens (Interview B.10).

When asked about deskilling of civil servants almost all interviewees agree that this will not become problematic. Some do acknowledge the risk of losing some skills and knowledge from the civil servant when the decision-making process is going to be partially automated to hybrid human-AI cooperation. However, this is not significant or is expected to be replaced with new skills and more time for citizens. There is however some fear that the public domain will cut on personnel after AI is able to overtake certain decision-making tasks of the civil servants one interviewee mentioned the Dutch tax authorities as an example where after implementation of an algorithm, a lot of civil servants were fired (Interview B.2). A few years later the system didn't operate as expected anymore, and all this missing knowledge and expertise of the civil servants had been gone. Others also mention the new skills that will come when civil servants will have to work together with algorithms, such as AI, in decision-making. A recent example, that has been mentioned in interview B.8, is the introduction of ChatGPT and the feared loss of writing skills of students in primary and secondary school. The interviewee acknowledged that this might occur but it could just as well result in an advanced reading skill where students become more critical in reading text. Authenticity will become more important. The most important part here is that AI should not overtake humans but, as mentioned before, *complement* humans. Having the best of both worlds will result in better decision-making and in more responsible AI implementation, where the civil servant may end up with an additional skill (Interview B.9).

Other remarks on the criteria for selecting a level of automation from the interviewees, that did not result from the desk research, are listed and briefly explained below.

1. *Impact on the citizens*: How does the level of automation impact the decision that is being made and thus impact the individual citizen?
2. *Technological capacity*: What level of automation is technologically possible in a certain scenario, also regarding the capacity of the system and the data?
3. *What is the goal?*: Why did we want automation? Why would this be beneficial?

These remarks are considered more to be context factors or to be points of attention that the policymaker needs to keep in mind, before even thinking about the criteria or the wanted outcome for the criteria when determining a level of automation for a certain decision-making task. Impact on citizens was mentioned by almost all the interviewees as the most important factor or criteria to take into account when deciding on a level of automation. What is the impact of choosing this level of automation, on this specific part of this particular decision-making process, on the individual citizen? Acceptability was mainly mentioned as a criterion for choosing a responsible level of automation when regarding the acceptability of citizens. This was considered very important together with the impact on the citizens, as even the most easy-to-automate tasks will not get high acceptability from citizens if any wrong decision turns out very hurtful for vulnerable citizens. This aspect has a lot to do with trust from the citizens which in the Netherlands has taken a blow after the 'Toeslagenaffaire', but may also be resolved partially after the positively received ChatGPT by most individuals, witnessing its potential. Next to this a more technological question was raised in some interviews with the scientific and the public domain experts, on the technological limitations of implementing a certain level of AI. This is also partially included in the set of criteria resulting from Chapter 4, where implementation effort is mentioned. The final remark is to step back from the whole decision-making process and ask yourself what the goal is in choosing a level of automation in this task. Sometimes human errors may be wanted because they can teach the civil servants or maybe you don't want to reject everyone that needs to reject according to the model because you want to keep a sense of humanity and dignity with the decision-making processes in the public domain (Interview B.6 & Interview B.7). As mentioned before a clear description of the goal of choosing a certain level of automation is needed in order to be able to implement it correctly and safely. One interviewee stresses considering the goal of automating a decision task and taking into account the impact and ripple effect that this level of automation may have on individual citizens but also broader on the things you cannot directly oversee, which also has to do with context factors as explained in the next section (Interview B.8).

Meaningful Human Control

The criterion on human control is discussed separately as the desk research showed that this may be considered the main criterion in choosing a responsible level of automation. This again resulted from the interviews with all the interviewees agreeing that in any level of automation, the civil servant should still be present in the decision-making process. Which immediately excludes again full automation without human control. From interview B.5 can be concluded that humans should always be present to be able to provide feedback and to change the model when it doesn't behave as expected. Following the Systems Safety Theory as mentioned in chapter 3 and as mentioned in Interview B.2, meaningful human control should include the civil servants being able to give input to the model and to impact the decision. This has been mentioned in multiple forms in all the expert interviews. The human should be considered in the decision-making loop (Interview B.9). We have to focus more on the role of the civil servant's presence in the automation of decision tasks and less on the algorithms (Interview B.2). As mentioned in Interview B.2 "*We have to focus on what the humans must be able to do in working together with algorithms to make good and responsible decisions*". Another interview elaborates on the importance of the civil servant to "*know, inspect, and give feedback to the model*. The interviewees were all asked to give aspects of meaningful human control, focused on the civil servant, these were categorised and coded in the following table. In this table, CS is used as an abbreviation for civil servant.

Table 5.3: Criteria for Meaningful Human Control, Coded Responses

Category	Public	Private	Scientific	Total
CS should have knowledge on the system	2	3	2	7
CS should have influence on the system	1	3	3	7
CS should observe the system	2	2	2	6
CS should be trained	2	3	1	6
Feeling of ownership	3	1	2	6
CS should trust the system	0	1	2	3
CS should have a clear goal	2	0	1	3
CS must be able to explain outcome	1	1	1	3

Implementing a level of automation where the civil servant has to work together with the AI to come to a decision or where the AI supports the civil servant, requires the civil servant to be taken along in every step of this policy implementation (Interview B.7). The civil servants will have to understand how the AI can help them, how it changes their role, what their responsibility is, how they should use it, and how they could potentially learn from it. On a decision level, you need someone who is equipped to understand what he or she is taking a decision about and for who (Interview B.9). When they are involved in the entire process of implementing AI as a decision support tool the acceptance of the civil servants will automatically be a lot higher. A balance needs to be reached in the tasks that a civil servant is going to do, in order to stimulate them positively as much as possible (Interview B.3 & Interview B.8). You do not want the civil servants to become brainless machines when the AI is deciding most for them and they only have to press the button. You will have to trigger them to think about the decision by combining simpler and more complex tasks. You could also have the human still take the decision and only use the AI as a control mechanism where the cases that are flagged by the AI get a revision (Interview B.6). The balance of the workload of the civil servants is important and as can be concluded from the interviews, it is not expected that the workload will decrease for the civil servants. *"Look at the past, if any capacity is available this is immediately taken through up-scaling of processes and by doing more and doing it faster."* (Interview B.9). But in order for the implementation of human-AI cooperation in decision-making, the civil servant needs to have enough incentives to stay focused and involved in the decision-making processes that directly impact individual citizens. The meaningful civil servant should, according to the interviews, have the following characteristics as presented in the list below. With a system or model, the implementation of a level of automation is meant, where the AI works together in a hybrid manner with the civil servant.

- The civil servant should have knowledge of the implemented system or model.
  - interviewees also mention the need for extra training or education for civil servants in order to achieve a needed level of automation.
  - Another interviewee mentions that using AI as a mirror for the civil servant can also help in creating more insights and knowledge for the civil servants
- The civil servant should be able to intervene and have an impact on the system that is implemented.
- The civil servant should be able to observe the system.
- The civil servant should have a clear goal on how to work with the system and what to establish together with the system.
- An aspect that is underlying most of these aspects is the level of trust by the civil servants in the system and in the way they operate together with the AI.
- All the points above will help the civil servant have a feeling of responsibility over the system.

5.1.3 Context Factors

Context factors are considered to be variable per case or per scenario where a decision is being made for an individual citizen. There are obviously also differences in decisions on the lives of individual citizens or the decision on how a certain document is stored. However, the scope of this research focuses on the cases where a responsible level of automation has to be determined for decision-making processes that have a direct impact on individual citizens. The interviewees were asked to give context factors that could differ within this scope. What different kinds of decisions can one think of and how can context factors differ where they should be considered when choosing a level of automation? As stated in interview 1: *"You can only provide an algorithm with the data that you have, and quite often that doesn't involve the context"*. Aspects of this context are for example the impact that it can have on individual citizens. This context factor (also mentioned as a criterion in most interviews) is mentioned in almost all interviews as the most important context factor that should have an influence on the level of automation. This is important because even for the smallest of decisions that seem so easily automated, the impact of a wrongfully taken decision can still be very big on an individual level as was also mentioned by Wirtz et al. [2019] (Interview B.8). The system will not be perfect and you have to think about the worst-case scenario. Another context factor that has been mentioned in multiple interviews is the sensitivity of the decision, the higher the sensitivity of a decision the more a civil servant would be wanted to interpret the context of the decision as well (Interviews B.2, B.3, B.4, B.6 & B.7). However, in one interview the expert argued that the sensitive decisions could also ask for a higher level of automation as this would reduce the emotional burden on the civil servant who doesn't have to decide on a personal matter anymore (Interview B.5). Although this expert also mentions that the AI needs to be very well trained on decisions made previously by civil servants and domain experts on those cases if you expect the AI to be able to overtake those decisions. In interview B.10 the key lies in using AI in a supportive manner where the civil servant takes the final decision and can question every aspect of the decision made by AI since this should be transparent. Another expert questions whether algorithms will ever come to a level where they would be able to copy the empathy of human beings and apply it successfully in decision-making (Interview B.6). Though questions could be asked on how ethical and how beneficial this is.

Another context factor that is also mentioned quite often is the repetitiveness of the decision task, as interviewees point out that the more often a decision occurs the more efficiency gain there could be when implementing an algorithm here (Interview B.2 & B.6). If a decision-making process is not repetitive enough there may not be a real incentive to automate this decision-making task. Another interviewee highlights that algorithms could also be used in identifying the more complex tasks which could also be beneficial for the recurrent decisions, in order to pay more attention to the more complex instances (Interviews B.4 & B.8). Saving time for the civil servant and thus increasing efficiency gain. This will only be beneficial when the decision is indeed recurrent. As was mentioned in the previous section, Repetitiveness is linked to the goal and purpose of using AI (Interview B.8. Do you automate because this increases efficiency or do you automate because this decision is too recurrent? As mentioned in interview B.8, if the latter is the case, maybe you should ask yourself why this decision has to be made more often as this could also be due to another problem where automating won't solve this problem.

Table 5.4: Context factors that play a role in determining the level of automation, Coded Responses

Category	Public	Private	Scientific	Total
Impact on Citizens	3	3	2	7
Sensitivity	1	3	3	7
Repetitiveness	3	1	2	6
Explainability	1	0	3	4
Knowledge (training)	0	2	2	4
Goal of using AI	1	2	1	4
Acceptability	1	2	0	3
Data Quality	2	1	0	3
Feedback loop for Citizens	1	1	1	3
Empathy	1	1	1	3
Changing External Factors	1	0	2	3

Other context factors that are also mentioned by some experts include the knowledge of civil servants in working with AI. They will need some training when they are expected to work together with an AI in decision-making, but it is also good to help them understand what happens behind the interface of an algorithm (Interview B.5 & B.6). What happens inside the algorithm? This is needed for them to be able to provide an explanation to the citizens when they ask how the civil servant together with the AI has gotten to this decision. This could be considered part of the transparency and explainability agreements that need to be made before implementing a level of automation. Because here, as with the concept of safety and equality, the definition of those concepts can be very ambiguous and thus needs clear mutual agreement. A number of interviewees have also mentioned the importance of defining a clear goal and purpose for using AI in decision-making. This may not be a clear context factor but is something that has been mentioned often and is needed in order to prevent wrongful or unsafe implementation of AI (Interview B.8). Especially when implementing AI from an external private organisation you need to make clear to the private party what is wanted from the public perspective, this requires a clear goal and understanding of what you are about to implement (Interview B.7. A final note that is worth mentioning is the changing external factors and environment in which a decision is made, which is mentioned as a context factor (Interview B.3). Changing situations in the world, with extreme cases such as an economic crisis and war, change our beliefs and our values and thus the context in which decisions are being made. As mentioned in Interview B.8, it could be the case that civil servants are more likely to follow the judgement of an AI during an economic crisis in fear of losing their job. An algorithm may be poor in noticing these environmental, more global shifts (for now), so human control and interpretation are, therefore, a must. The level of automation should maybe also be changeable over time since the data used in an algorithm could be outdated. An easily adjustable level of automation can result in more flexible policy adjustment and quicker advancements when the technology significantly improves from an ethical viewpoint.

5.1.4 Responsibility and Autonomy

Since human control was mentioned by all the interviewees as an important criterion for determining the level of automation, as a result of the desk research as well, it is helpful to define what the role of the civil servant should be in a hybrid system where human and AI cooperate in the decision-making process. The interviewees were asked what the role of the civil servant should be and how this is linked with meaningful human control. Following the requirements for meaningful human control as established based on the interviews, the role of the civil servant can already be given form to a certain extent. This section will dive more into the responsibility and autonomy aspect of human control and of the role of the civil servant.

Based on the interviews, the civil servant should already be present in the design phase of the system where they can already explain their expertise and elaborate on how the algorithm can support them and how it should operate in the decision-making process (interview B.2). This will help in creating a feeling of ownership as it increases the acceptability of the civil servant regarding the usage of such hybrid human-AI systems in decision-making. It was also mentioned that acceptability should not be regarded as a form of acceptance as it implies that something has happened to them (Interview B.8). It is advised to speak of a level of ownership and authority over the decision and the hybrid human-AI system instead of acceptability by the civil servant. You want to include the critical thinking of the civil servants when implementing AI and thus there needs to be a clear role for the civil servant in the decision-making process. And again, this feeling of ownership and more authority also aids in increasing the trust of civil servants in these systems which will benefit in creating meaningful human control. Involving them in the design phase will also aid in their understanding and their ways of communicating with the AI. The communication part is seen as a very important aspect of the role of the civil servant in such hybrid systems (Interview B.4). As soon as the civil servant knows the system, is able to communicate with it and interferes whenever that may be needed, this could result in more transparency and explainability. One needs to understand how the AI has aided to get to a certain decision together with the civil servant. As mentioned during Interview B.3 "If you don't understand how something got somewhere, by definition if you don't understand how something works, chances are that malfunction will occur at some point. You won't be happy because you don't know how it works and where it has gone wrong." In the case of public decision-making, we need to be able to explain how a hybrid system, where AI and civil servants work together, comes to a decision. If the AI has done the data analysis, it needs to be clear, how it reads the data, how it uses the data, how it cleans the data, what it adds or deletes to the data, how it transforms the data etc. If it advises on taking a certain decision, it needs to be clear how the AI has gotten to this advice, how the AI analysed the data to get to this decision and how the AI traded off certain criteria and variables.

Next to this, the civil servant needs a level of awareness of how the usage of AI can have an impact on the way they make decisions (Interview B.5 & B.10). Civil servants may become oblivious to the fact that they have been following the judgement or advice of the AI too soon without really being triggered to think it through thoroughly and to ask questions about the advice that the AI has given. For this, it is important that the civil servants know how they can ask those questions and how they can interfere with the system. They have to know "what buttons they need to press in order to change or adapt a decision" (Interview B.7 & B.9). And next to this elaborate on why they disagree, this triggers the civil servants to think thoroughly and consciously about their decision, the role of the AI and how this has an impact on the citizens (Interviews B.4, B.7, B.9 & B.10). The civil servant eventually "needs to be the critical link" in the chain of this decision-making process, becoming the "glue in between organizational parts" where they can explain what is happening and why it is happening (Interview B.8).

A side note needs to be made on this constant triggering of the civil servant, as they also want to feel progress. It is not needed for the civil servant to critically reflect on every piece of advice given or task done by the AI. This will result in a more inefficient system than is currently in place. "Preferably a selection is already made to filter the less complex cases and have those cases done with a lower level of human control than the more complex ones" (Interview B.9).

Responsibility

Responsibility is an enormous challenge in using AI, especially in decision-making processes in the public domain where the outcome directly impacts individual citizens. It is not clear-cut who is responsible in the case where a wrong decision is being made. Multiple actors are present in this implementation of AI, such as the governmental institution, the civil servant or operator of the AI, the manager or policy maker who decides on the implementation of such systems, and also the creator and provider of the AI (when this is developed outside of the government organisation). Almost all interviewees agree on the shared responsibility in such hybrid human-AI systems, dividing responsibility over multiple involved actors. The interviewees also mention that the civil servant could be partially responsible as they could have intervened when a wrong decision was being made working together with the AI. Provided that this is possible. The policy maker who decided on the usage of this hybrid system should be partially (if not fully (Interview B.3)) responsible. In most interviews, the main responsible actor should be the one who decided upon the implementation of AI in a certain decision-making process (Interviews B.3, B.6 & B.8). The other interviews mention slight deviations from this with different roles for the civil servant. It is expected by interviewees from the public domain, that the civil servants feel responsible for the final decision (Interview B.4 & Interview B.10) and not become too lazy where they blindly follow the algorithm's advice and stay critical of the outcomes of the algorithm, as mentioned by experts from the private domain (Interview B.5, B.6 & B.7). An expert from the public domain stresses the importance of looking at this development as a huge change in how public organisations now run and organise their decision-making processes (Interview B.4 & B.9). The civil servant could become the glue within an organisation, balancing the need and the requests of the citizens working together with AI and implementing but also questioning the policy provided from higher up in the organisational hierarchy (Interview B.8). It should become clear that civil servants have the most field knowledge and are in close communication and involvement with the citizens. They need to use this knowledge and expertise to question and improve the policies and rules that are provided by managers and policy-makers (Interview B.4). This is the responsibility of the civil servant, also regarding the implementation of AI, with a certain level of automation, in the decision-making processes by civil servants, summed up in the list below. The civil servant:

- is expected to be critical of the decision-making process and the implementation of an algorithm in this process.
- is expected to critically use the outcomes of the algorithm, and critically reflect on their own decision-making and patterns that appear through the use of an algorithm.
- is expected to be aware of the impact of their decision-making (whether in cooperation with an algorithm) on the citizen.
- should help and debate with other colleagues when working with algorithms in order to reach a level of equality in the way civil servants use the implementation of an algorithm.

As mentioned in interview B.9, the civil servant nor the AI can be ultimately responsible in case wrong decisions are being made or when something goes wrong regarding biases, discrimination, exclusion, etc. They both have a certain responsibility but to an extent. The policy maker who decides on implementing the AI is ultimately responsible, in the end, this could be the board of directors or the Minister. The figure 5.1 below shows how responsibility should be divided among the three actors involved based on the results and discussions from the interviews. The three different domains mostly agreed on this shared responsibility with one ultimately responsible actor.

5.1.5 Interview Conclusion

The interviews resulted in very valuable insights into the way experts look at the criteria, focusing on a civil servant perspective, that can be examined when deciding upon a level of automation for a specific decision-making task by civil servants. The context factors provide more nuance in deciding upon a responsible level of automation with regard to the impact that this level of automation may have on citizens. The final part focuses on the responsible side of choosing a level of automation for which different actors should be responsible which, when combined, results in a responsible implementation of algorithms in supporting civil servants in their decision-making processes. It also elaborates on the knowledge that is needed for policy-makers that they need in order to implement algorithms responsibly. More

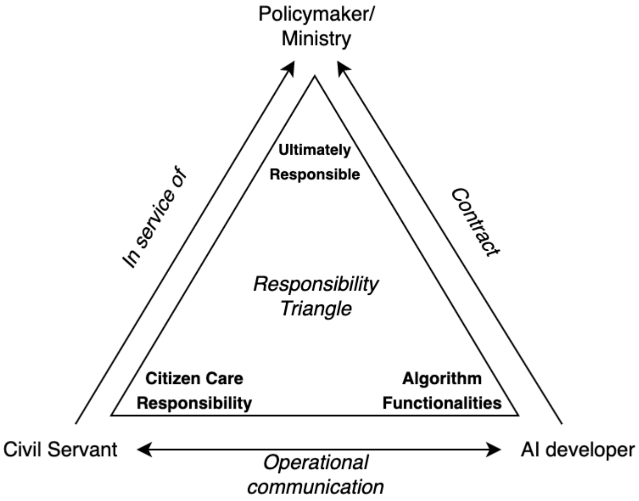


Figure 5.1: Responsibility Triangle with their relationships

on this later in this chapter.

Resulting from this section some main criteria were found that are needed for the next section. These are the quality of the decision and the efficiency of the decision, not to mention the importance of human control by the civil servant. By the interviewees, these are often traded off with the impact that this has on the individual citizens. The impact is therefore also considered together with the main context factor which is the sensitivity of the decision that is supported by a level of automation with the implementation of AI. The main criteria will be used in the following section together with the main context factors in order to create four vignettes where trade-offs can be measured in a vignette experiment. The setup of this experiment will be explained in the next section.

5.2 VIGNETTE EXPERIMENT

In vignette experiments, multiple factors and criteria can be tested by means of changing the values for those criteria in the same and different context [Atzmüller and Steiner, 2010]. By changing the context factors slightly as well, the impact of the context factors on these criteria and the way people look at them can also be investigated [Langer, 2016]. For this vignette survey experiment, two cases are created in which the decision-making process of civil servants in municipalities is supported by a level of automation through the implementation of AI. These two cases differ in their context factors, mainly sensitivity of the decision. For each of these cases, two scenarios are written out that differ in the main criteria that resulted from the expert interviews, mainly human control and efficiency gain. First, the group of respondents will be elaborated on, explaining how and why this is the selected group. Second, the two cases will be described before diving into the four scenarios that have been created. Finally, the results from this vignette experiment will be discussed showing the different perspectives between the vignettes but also trying to look at any differences regarding the three different domains as also discussed in the interview section (public, private and scientific). Figure 5.2 shows a conceptual diagram of how these vignettes differ over the input variables, which are the criteria and context factors. This may help in visualising how these vignettes differ from each other. It is also easy to see now that there is no vignette or scenario in the top right square, which helps to visualise that it is still very hard to implement an automation tool in a responsible manner so that there is a high increase in efficiency but also retain a high level of human control.

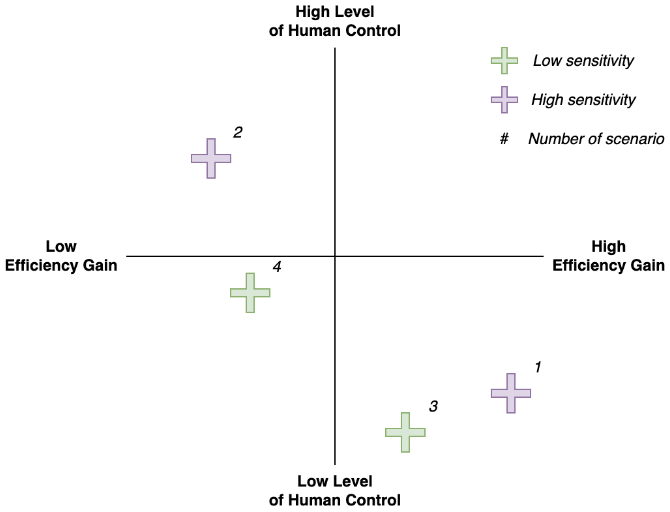


Figure 5.2: Vignettes on the scales of the criteria and context factor

5.2.1 Survey Sample

The respondents for this survey were all gathered at a congress on the responsible use of AI in the public domain, and specifically on the Dutch public domain. This results in a group of respondents that is very interesting from a qualitative point of view as they have more knowledge than average on the implementation of algorithms in the public domain. Most people present at this congress are working in the public domain though people are also present who work in the other two domains that were also investigated in the interviews: the private and the scientific domain. Though not all respondents work in the public domain they were asked to look at 4 different hybrid human-AI cooperation in decision-making in municipalities by civil servants. They were asked to look at this from a civil servant perspective. It is interesting to see if there are any big differences between these 3 domains and also within these domains. One thing the respondents have in common is an interest and a certain level of knowledge in the implementation of AI in the public domain.

40 respondents filled in the survey, of who 26 filled it in completely. This could be sufficient for qualitative research as it gives some first insights into how people working in this domain view the issues and trade-offs at hand, it is however impossible to say something about the larger population. Therefore the results of this vignette experiment will most likely not be significant though they can provide us with valuable insights and function as a starting point for more research into determining a responsible level of automation in different domains and tasks. Also because this group of respondents is quite a niche with a high level of prior knowledge and a high association with this topic and domain. Because the reasons behind some of the choices of these respondents may be very interesting to know, they were all given the possibility to explain their choice after rating each vignette. This was completely voluntary and could help in understanding the challenges and the trade-offs that they see. Due to the smaller group of respondents and the qualitative nature of this research, no randomisation was applied to the way the vignettes were presented to the respondents [Atzmüller and Steiner, 2010; Heverly et al., 1984]. This makes it easier to analyse the open questions and the reasoning behind the choices made for each scenario by all the respondents.

5.2.2 Case Description

The vignette experiment is divided into two cases with each two scenarios or vignettes. The two cases mostly differ in the sensitivity of the decision that is automated through the implementation of AI which supports the civil servant. This automation is defined per vignette and will be explained in the next paragraph. The two cases have been chosen based on personal communication and interviews with experts from the public domain. The main requirements were that they had to differ in sensitivity but also had to be processes that are part of the same public organisation. For example not looking at comparing municipalities and immigration or tax services as these have different organisational components that may impact the outcome of the survey. For the two cases, the decision-making process of civil servants working in municipalities has been chosen. The two cases had also been chosen in order to be easily understood and with a clear difference in their context factors. It had to be easily understood for the people filling in the survey what the task of the civil servant would be, what the impact on the citizens could be and how AI could support this particular decision-making process. Both cases, therefore, need to be very clear, interesting and easy to comprehend for the respondents.

The first case is the more sensitive decision-making task and entails the public service of providing special welfare to those who are in need of this. Whether people are entitled to these subsidies depends on a number of criteria that deal with sensitive data. It basically entails that when you don't have a sufficient income and face necessary costs, this can be covered by the municipality. This can be either a monthly contribution or a one-time financial aid by the municipality [Rijksoverheid, 2023b]. Since this special regulation is mainly meant for those who are most likely to have some trouble making ends meet, this is very sensitive. Especially for citizens living in harder and tougher living conditions. The cases are explained below (in Dutch, as in the survey, with English translation) and include 2 scenarios each that will constitute the four vignettes as will be explained in the next section.

Vignettes

The two cases as presented in the previous section each have two scenarios that differ in the main context factor on the sensitivity of the decision and the criteria that focus on efficiency and human control of the human operator, in this case, the civil servant. They are asked about the perspective of the civil servant, though obviously, not all participants are working in the public domain. It is however also interesting to see how the other two domains look at these issues and whether they are on the same page or not. All of the respondents do at least have an interest in AI in the public domain in common, whether that is AI expertise, from a scientific point of view, maybe as the provider of an AI tool or just merely because they want to learn more. Due to the limited number of vignettes, criteria and context factors that are investigated, no randomisation has been adapted as this is more in line with the classical vignette designs, as mentioned by Atzmüller and Steiner [2010]. The respondents are asked to rate each vignette on a scale from 1 to 5 on how willing they would be to work with such a hybrid human-AI system in the decision-making of civil servants. 1 meaning *absolutely not* willing to use such a system and 5 meaning they would be *absolutely* willing to use such a system.
As with the cases, the scenarios were also presented in Dutch to the respondents as this was the mother tongue of most of them, while the rest mastered the Dutch language well enough. The translation of those scenarios is placed below while the original scenario sketches are presented in appendix C, together with the demographic question and the open questions. For each of these scenarios, the respondent had to rate the scenario on a scale (Likert (Ordinal)). Scenarios 1 (*SW Efficient*) and 2 (*SW Control*) are described for the first case (Special Welfare) and scenarios 3 and 4 (*PP Efficient* and *PP Control*) for the second case (Parking Permit). The

Table 5.5: Case descriptions as provided in the vignette experiment (in Dutch with translation)

Casus 1: Bijzondere Bijstand
Ambtenaren beslissen, op basis van een aantal criteria, of een burger recht heeft op de zogeheten Bijzondere Bijstand (link naar Rijksoverheid). Deze beslissing hangt af van gevoelige data zoals de financiële of gezinstoestand van een individu. Hieronder worden tweescenario's geschetst waarin een algoritme samen met de ambtenaar deze beslissing neemt. Dit is een hybride vorm waarin de mate van controle verschilt voor de ambtenaar en algoritme voor de verschillende scenario's.
English: Special Welfare
Civil servants decide, based on a number of criteria, whether a citizen is entitled to the so-called "Bijzondere Bijstand" (special welfare) [Rijksoverheid, 2023b]. This decision depends on sensitive data for example the financial or family situation of an individual. Two scenarios will be described below in which an algorithm has been implemented in the decision-making process where it supports the civil servant in coming to this decision. This is a hybrid form in which the level of control differs for the algorithm and for civil servants in different scenarios.
Casus 2: Parkeervergunning .
Ambtenaren beslissen, op basis van een aantal criteria, of een burger recht heeft op een parkeervergunning (link naar Rijksoverheid). Deze beslissing hangt af van een aantal voorwaarden zoals een kenteken dat op naam van de betreffende burger moet staan. Dit kan per gemeente verschillen. Voor nu gaan we ervan uit dat het niet gevoelige data is welke gebruikt wordt (woont de burger in dit gebied en is het inderdaad de (enige) auto van de burger). Hieronder worden twee scenario's geschetst waarin een algoritme samen met de ambtenaar deze beslissing neemt. Dit is een hybride vorm waarin de mate van controle verschilt voor de ambtenaar en algoritme voor de verschillende scenario's.
English: Parking Permit
Civil servants decide, based on a number of criteria, whether a citizen is entitled to a parking permit [Rijksoverheid, 2023a]. This decision depends on a number of criteria such as a licence plate that has to be registered to the name of the corresponding citizen. Some criteria may differ slightly per municipality. For now, it is assumed that no sensitive data will be used (the citizen should be living in the corresponding area and this car is indeed their only car). Two scenarios will be described below in which an algorithm has been implemented in the decision-making process where it supports the civil servant in coming to this decision. This is a hybrid form in which the level of control differs for the algorithm and for civil servants in different scenarios.

vignettes are used to identify developments and data on sensitive, ethical issues, where attitudes and beliefs on the responsible implementation of AI regarding the trade-off between efficiency gain and human control by civil servants are mapped.

Each of the scenarios differs in efficiency gain, which is noted as time efficiency gain in these examples. Human control is the second criterion that differs where in certain scenarios human control is described as the way in which civil servants have control over the final decision and the way the algorithm acts. The context factor sensitivity is taken into account by providing two different cases which both use different kinds of data to come to a decision, namely more sensitive data in case SW and significantly less sensitive data in case PP. The criterion of quality is also slightly implemented as one could say that providing more explanation and more attention to the citizens results in a higher quality of the decision or maybe even more acceptability by the citizens in such a system. This is due to the fact that some of these criteria will slightly overlap. This could, on the other side, also be seen as efficiency gain as the civil servant will have more time to efficiently and more meaningfully provide an explanation of a decision. The results of this vignette experiment will be analysed below, where the trade-offs between these criteria and context factors will be shown.

Table 5.6: scenario descriptions as provided in the vignette experiment (*translated*)

Case 1: Special Welfare (SW)
Scenario 1, name: SW Efficient
By making use of the algorithm the decision can be taken within 5 minutes, which normally takes about an hour. This enables civil servants to spend more time in direct contact with citizens having more time to explain certain decisions that have been made (more attention for the human). The algorithm here advises a certain decision to the civil servant. There is little control for the civil servant as they base their choice on the advice given by the algorithm. Next to this, it is unclear how the algorithm got to this advice and how certain criteria have been traded off by the algorithm. The result is that the civil servant has little impact on the making of the decision.
Scenario 2, name: SW Control
By making use of the algorithm, the data, needed to come to a decision, is directly analysed and clearly presented in an overview for the civil servant. This results in the civil servant being able to take a well-informed decision within 45 minutes instead of 1 hour, which it normally takes. The civil servant thus has more time to explain the decision to the citizens. How the algorithm analyses the data is comprehensible and accessible for the civil servant. The civil servant here has the possibility to intervene and to have another look at the data when they have doubts about the overview provided by the algorithm.

Case 2: Parking Permit (PP)
Scenario 3, name: PP Efficient
By making use of the algorithm the civil servant can grant or decline a parking permit within 5 minutes. This enables civil servants to spend more time on personal attention with more complex decisions. The algorithm makes a decision on issuing a parking permit based on a number of criteria and the civil servant's only task is to implement this decision or not. How the algorithm comes to a decision is not known but if there appears a case that really deviates from what normally happens, the civil servant has the ability to intervene.
Scenario 4, name: PP Control
By making use of the algorithm the civil servant can grant or decline a parking permit within 15 minutes. This enables the civil servant to spend more time providing an explanation for this decision. The algorithm gives a clear overview of the criteria and provides the civil servant with advice. The civil servant can deviate from this advice when this is considered not the right decision in the eyes of the civil servant. How the algorithm comes to this advice is unknown.

5.2.3 Results

This section shows the results of the vignette experiment first by looking at the differences between all the respondents for the different vignettes. What criteria do they seem to value higher over the other and how does this change in another context? What trade-offs do they make between efficiency and human control? Second, results can also be shown by looking at the differences between the three domains. This is the only demographic that is used in this experiment as the main goal is to look at the reasoning behind the trade-offs and the ratings. It could be interesting to look at the differences between these domains and to point towards interesting topics of discussion that need to be held in order to get an agreement on all the different values and concepts that need clear definitions. Last but not least, the results are used to see how looking at these trade-offs of criteria and the impact of certain context factors can help in determining a (more responsible) level of automation that can support the civil servant in their decision-making. Before diving into the trade-offs and the differences between the domains, the general results are shown below to look at the scenario that has gained the highest rating, which means that most respondents were relatively positive about using this hybrid model of human-AI decision-making.

It has to be noted that this is far too short-sighted to tie an undeniable conclusion to these results as the group of respondents to this vignette experiment was too small, and no significance was found in any of the results. Nonetheless, this research aims to take the first steps in providing guidance for those willing to experiment with AI in the public domain. The results from the vignette experiment together with the interviews and desk research results in very valuable insights, reasoning and perceptions from different domains that give a starting point for determining a responsible level of automation, mainly focusing on the trade-off between efficiency gain and human control for different sensitive cases with a clear role for the civil servant.

Table 5.7: Descriptive statistics for the four scenarios

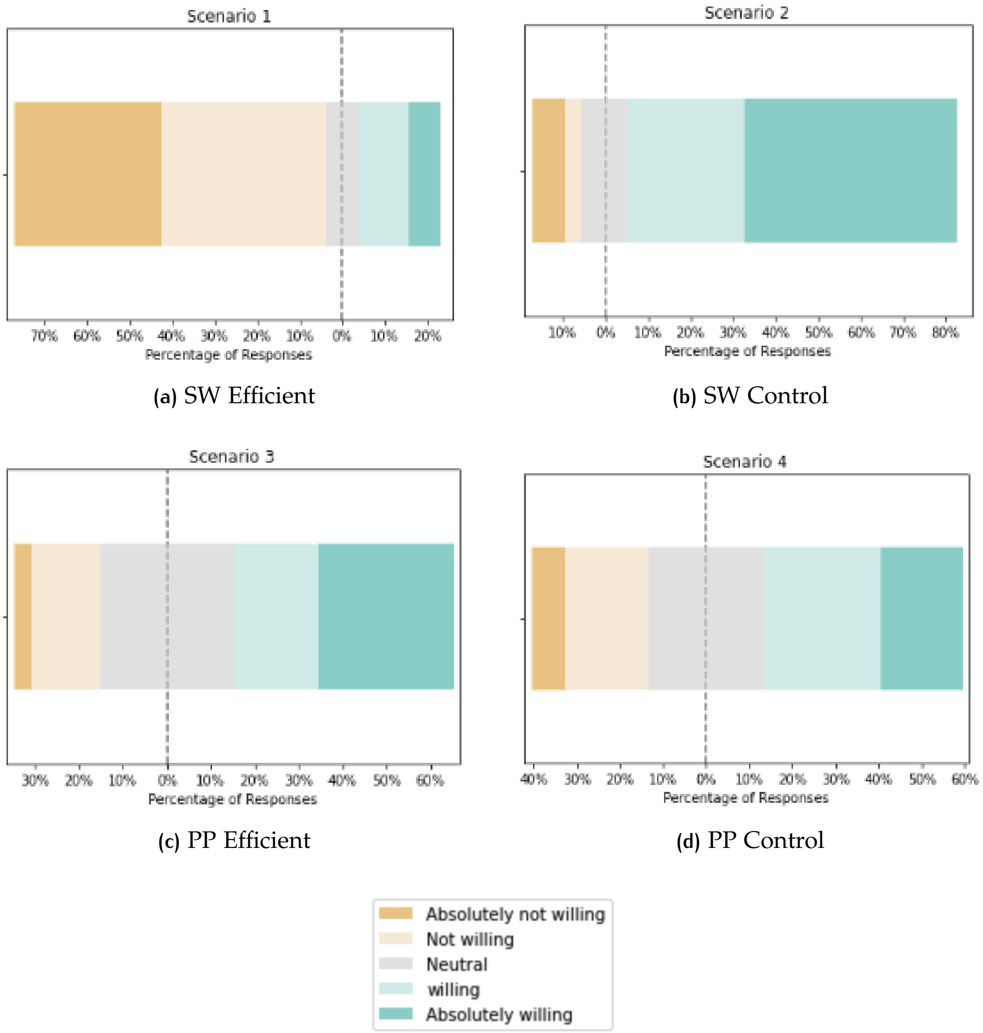
	Case SW		**Case PP**	
Descriptive Statistics	***SW Efficient***	***SW Control***	***PP Efficient***	***PP Control***
Mean	*2.19*	*4.08*	*3.58*	*3.31*
Standard Deviation	*1.27*	*1.23*	*1.21*	*1.23*
Median	*2.0*	*4.5*	*3.5*	*3.0*

From table 5.7 it is clear that Scenario *SW Control* is rated the highest on average which means that the human-AI system as presented in this scenario is the likeliest to be used by the respondents. This is the scenario where there is a slight increase in efficiency gain but foremost the civil servant is possible to intervene in the decision-making process and the algorithm is completely insightful which results in a high level of human control and also a high level of transparency, as the civil servant comprehends how the AI got to a certain decision. Looking at the differences in the two cases one can see that in between those cases, the averages differ for the scenarios in case SW but not for case PP. These are roughly the same even though scenario *PP Efficient* results in a lot more efficiency gain for a non-sensitive decision-making process. The level of human control differs between scenarios *PP Efficient* and *PP Control* in that the civil servant can only decide whether or not to implement the decision suggested by the algorithm in scenario *PP Efficient*. Whereas in scenario *PP Control*, the civil servant can also deviate from the advice given by the algorithm. For both scenarios in case PP, there is a low level of transparency which result in a lower level of human control as the civil servant doesn't completely understand what happens inside of the algorithm. Thus this could show that even though there is a difference in efficiency gain there is no high preference over the other scenario due to the lack of significant human control. Note here again that the results are not significant, so they cannot be interpreted as the undeniable truth though they do give us some insights into the beliefs and attitudes of the respondents.

Figure 5.3 shows the different distributions for the four different vignettes or scenarios on how willing the respondents are to use these kinds of hybrid human-AI systems. Scenario *SW Efficient* can be seen as the least favourite scenario for the respondents where there is a high-efficiency gain but a low level of human control for a very sensitive decision-making case. Scenario *SW Control* has the highest percentage of people *absolutely willing* to make use of such a system, which shows the opportunity for automation tools and a certain level of automation also for highly sensitive decision-making cases. Whether this level of automation as described in this scenario is responsible, will be discussed later showing what policymakers willing to experiment with these levels of automation need to consider and think about before taking deciding upon a level of automation.

Trade-Offs Criteria and Context Factors

Human control is considered to be the main criteria that respondents take into account, as can be concluded from the desk research and interviews. When looking at



(a) SW Efficient

(b) SW Control

(c) PP Efficient

(d) PP Control

Figure 5.3: Likert distributions for the four vignettes

the vignettes, it can be seen that as soon as there is a certain (high) level of human control, any efficiency gain is welcome and thus the hybrid human-AI system is most likely to be adopted, following the results from the survey. However, if this level of human control cannot be guaranteed, a significant increase in efficiency will not lead to a certain trade-off where a lower level of human control is quickly accepted. Especially not in sensitive decision-making cases. As soon as the decision is less sensitive and probably also has a lower impact on the citizen, this trade-off starts to show a little more as can be seen comparing scenarios *PP Efficient* and *PP Control*. In scenario *PP Control*, the civil servant has slightly more control as they can look at the advice provided by the algorithm and decide for themselves whether they agree with this or not. In scenario *PP Efficient* the only control the civil servant has is whether to implement the decision made by the algorithm or not. Interestingly, scenario *PP Efficient* has a slightly higher average (*3.58*), which could mean a higher willingness to implement such a hybrid human-AI system, compared to scenario *PP Control* (*3.31*. However, the difference between scenario *PP efficient* and *PP Control* is not significant ($p = 0.43$) whereas the difference between scenario *SW efficient* and *SW Control* is ($p = 0.0000016$), conducting two-sampled t-tests. This could be explained by the efficiency gain which is higher in scenario *PP Efficient* (< *5 minutes*) than in scenario *PP Control* (< *15 minutes*). This would suggest that for cases with a lower level of sensitivity, the level of human control is not as important as soon as the efficiency gain is significantly higher. As is mentioned by some of the respondents as an explanation for their reasoning ”it is less sensitive” but again ”not enough efficiency gain”. Thus in cases with a lower level of sensitivity, the hybrid human-AI system with a higher efficiency gain is more likely to be used by the respondents than with a lower level of efficiency gain. Although the majority says that there is a need for explainability and transparency and therefore the average of both systems lies slightly, but barely, on the positive side of the willingness to use any of these two systems.

When comparing the two cases that differ in levels of sensitivity, and arguably also in the impact that any wrong decision may have on the individual citizen, it is interesting to see that as soon as a certain level of human control is present, the sensitivity does not seem to be too much of an issue anymore. Thus one could conclude that a civil servant’s discretion and ability to consider the context is considered to be present and positively impacts the decision that is made. This is in line with the remarks made by Mitrou et al. [2021] where the discretion of civil servants is needed to be able to indicate what the impact of these decisions and of the hybrid human-AI system is. When looking at figure 5.3, at the sub-figures 5.3c & 5.3d, it can be seen that a large percentage is neutral on whether they are willing to use such a system whereas this non-sensitive decision could also be considered to be easy to automate and not with great risk. Which would result in a higher willingness to use these systems. This is also mentioned by some respondents who did give this vignette a higher score where they mention that with those easily automated processes, a lot of steps could and should be automated. This could be explained by saying that some respondents, consider the possibility of automation, even with a low-efficiency gain, a reason for automation on its own. This was not however stated so bluntly by the respondents.

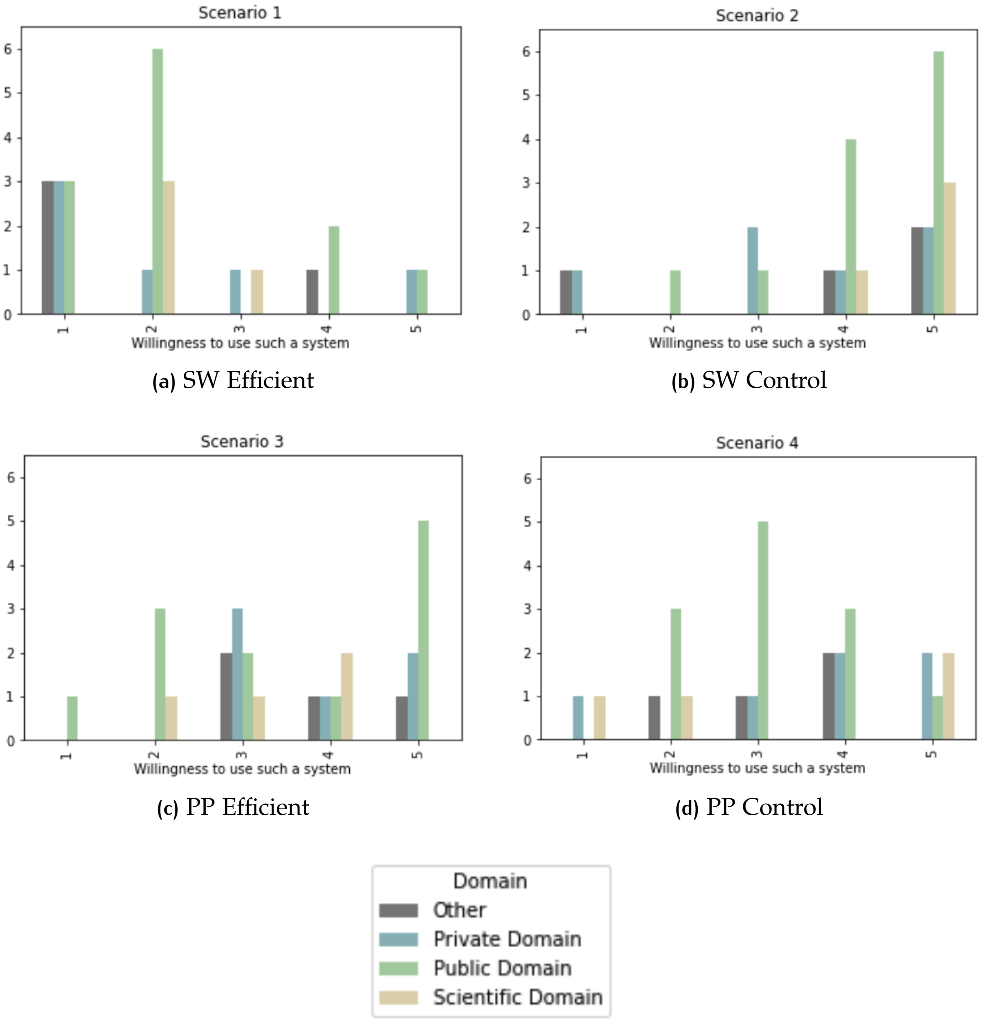
The survey experiment also revealed that the respondents consider different levels of responsibility in the scenarios presented, as was also concluded from the interviews. For instance, in case PP, one respondent believed that it is the citizen’s responsibility to notify the civil servant and the municipality of any issues with the parking permit. While another respondent notes that the civil servant should be knowledgeable about the system and able to provide an explanation. This suggests a need for increased human control in such cases. While some respondents preferred a lower level of human control in the less sensitive scenarios, others cautioned that explainability is still important to ensure controllability. These findings highlight the complexity of assigning responsibility in certain situations and underscore the importance of considering multiple perspectives when designing such hybrid decision-making systems.

Interestingly, the interviews also mentioned that decisions such as a parking permit, with a lower level of sensitivity, will most likely require a lower level of human control where more efficiency gain is preferred. One interviewee however notes that even though this decision may seem to have less impact due to the sensitivity of the decision, for an individual citizen under certain circumstances, such a decision can still have a very high impact with huge consequences, not to mention on the mental health of the citizen.

Domain Differences

When looking at the different domains that have filled in the survey we can see that most respondents are working in the public domain (12 respondents) as was also expected when considering the goal of the survey (civil servant perception) and the Congress where the respondents were asked to fill in the survey. The private domain counts 6 respondents and the scientific domain counts 4 respondents with the remaining 4 respondents working in a combination of those domains. One could expect that when looking at the differences between the three main domains, the private domain (of whom mostly startups working with an AI tool) would have a higher willingness to use any AI system. However, this is not seen in the data, as can be seen in figure 5.4. This could be due to the low number of respondents or because of the awareness of the issues that arise with implementing AI in the public domain as they may have experienced when working with public organisations as mentioned in interview B.6. One of the respondents working in the private domain also elaborates on the main issue of transparency, where the first scenario is an absolute no go but the other three seem to be okay as long as the decision stays with the civil servant or when there is the possibility to intervene in less sensitive cases.

When comparing the scores for each scenario in the three different domains, another interesting pattern arises, this can be seen best in the visualisation in figure 5.5. For the first two scenarios, which are the scenarios with a higher sensitivity and possibly a higher impact on the citizens, the scientific domain is the most positive about both scenarios with a clear distinction for scenario *SW Control*. Though no significance was found here, due to the limited amount of responses, it also came to light in interviews B.6 and B.7, that the private domain is slightly more hesitant in more sensitive cases. They mentioned the technical restrictions to include context in AI systems and the low level of acceptance from citizens in automating more sensitive decisions. Comparing the private and scientific domains in scenario *SW Control*, the difference between their averages is the largest however no significance has been found ($p = 0.107$). For the third and fourth scenarios, the domains seem to have flipped their attitude towards using such systems. For these scenarios with a lower sensitivity and possibly a lower impact on the citizens, the scientific domain seems to be the least optimistic about using such systems whereas the private domain sees more opportunities. The public domain is sort of meeting in the middle of these other two domains. This could obviously be the case due to the low number of respondents and thus further research would be necessary. No significance has been found in comparing any of the domain averages for each vignette. It is however interesting to see and when looking at the explanations given by the respondents, at least some points worth noting can be found regarding the trade-off between human control and efficiency gain. One respondent from the private domain mentions that especially in sensitive cases, transparency of the entire decision-making process is key for the civil servant in order to validate whether the



(a) SW Efficient

(b) SW Control

(c) PP Efficient

(d) PP Control

Figure 5.4: Distributions per scenario for the different domains

decision is based on the correct and valid arguments and data. This is less the case for the less sensitive decisions, as long as the civil servant still has the possibility to intervene. The scientific domain focuses more on explainability where the sensitivity of the decision is less of a relevant factor. As long as the decision-making by the algorithm can be explained and is transparent, any efficiency gain is welcome and both the citizen and the civil servant can benefit from such a system. A side note is the possibility for citizens to always have the ability to raise an alarm when a wrongful decision or when an error has been made.

One respondent mentions the influence that biases can have on the way an AI algorithm performs, specifically in sensitive cases. The bias that is spoken of could come from the maker of the AI or from the data that is the input of the AI which creates a higher risk of discriminating against certain groups of people. With a low level of human control, there is no room for a tailor-made approach. Some respondents see the same risk for both cases (sensitive and not sensitive) which was also mentioned in interview B.6, where wrongfully (not) giving out a parking permit may not seem so harmful, but for an individual with for example a handicap this can have a very high impact. Another respondent does not see an issue at all with any of the scenarios as long as the selection and implementation of the algorithm have been done carefully. This respondent also mentions the risk of too much ambiguity and discretion by civil servants. If every civil servant can just deviate from the advice given by the AI, why implement an AI at all? These responses could be seen as extremes though it is useful to note them and take them into account. The trade-off that most respondents make in the first scenario is that, due to the high level of sensitivity, a high level of human control with a transparent and controllable system is mandatory. If this is not present, even if the efficiency gain is super high, they do not prefer to use the system when it replaces the old system. If it runs next to the old system and the civil servant can make use of both in comparing and learning from these systems this could be a solution for implementing less- or non-transparent systems. Different views also exist on using transparent systems that still dependent on a certain set of historical data which could be biased, where the civil servant does not have the training and skill to work with an AI and perform their role as a ”human-in-the-loop”.

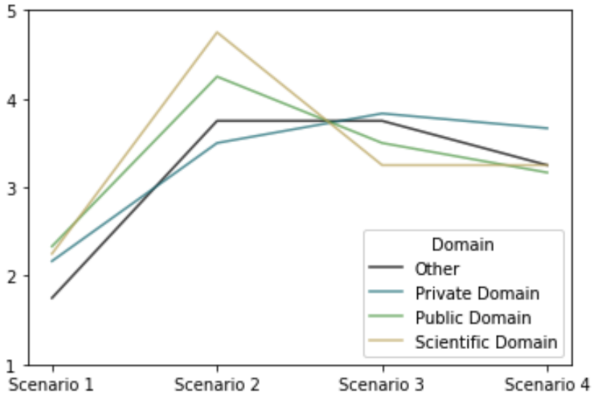


Figure 5.5: Domain Averages

For scenario *SW Control* the efficiency gain is not super high where a decision can be made within 45 minutes when using the hybrid human-AI system vs. the current situation where it can be done within an hour. These numbers have been made up, but help in visualising also for respondents who are not familiar with the public domain to see what kind of efficiency gain automation could bring. But since the civil servant can have complete control over the decision-making process and every step that the AI takes is insightful, most respondents have a high willingness to use this system, as can be seen in figure 5.4b. This also builds on the results of the interview and the desk research that as long as the level of human control in highly sensitive cases is sufficient, any improvement in efficiency is welcome. The other way around does not work as the civil servant needs to be able to argue and question the algorithm, this is not possible if the algorithm cannot explain or reason why it has taken a certain decision, as noticed by a respondent from the public domain. Testing and experimenting with these hybrid systems are mentioned by multiple respondents as a way of making them more reliable and responsible. It has to gain trust, even if it is not transparent and thus no room for an explanation by the civil servant, according to one respondent.

Level of Automation

These findings above can provide guidance to those thinking of implementing a certain algorithm, such as AI, choosing a responsible level of automation. As results from chapter 4 multiple facets need to be taken into account in order to implement a decision support algorithm such as AI in order to be supportive for the civil servant and to the citizens. This research has focused on the mapping, analysing and testing of the criteria that determine the preferred and responsible level of automation by experts working in the three domains that work with AI implementation in the public domain, with a focus on the perspective of the citizen and what the human control should be, regarding the role of the civil servant. In order to give advice on how to use these trade-offs and this analysis the following suggestions are made.

Based on chapter 4, four implementations are established in table 5.8 below. These are based on the findings from the interviews and the literature study, and give some insights into how different values for the criteria can result in a certain level of automation that may be preferred or more responsible. The results from the survey will be used to determine a responsible level of automation from these four implementations, based on the criteria trade-offs and the context factor analysed. The level of ”No automation” and the level of ”Full automation” are excluded from this research as explained in chapter 4, as the trade-offs in the hybrid human-AI systems are the main target for this research. The hybrid levels of automation where humans and AI together take the decision are most important when looking at the trade-off between efficiency gain and human control from a civil servant’s perspective. The four implementations that have been sketched in table 5.8 follow the levels of automation as mentioned in table 4.3.

Based on the findings from the desk research, interviews and the trade-offs that resulted from the vignette experiment on efficiency gain, human control and sensitivity, a responsible level of automation for case SW, with a sensitive decision-making process, should have a high level of human control excluding levels such as routing and decision-support as mentioned in a simplified categorisation in table 5.8. Regarding the sensitivity of case SW, a lower level of automation may be preferred as the impact on the citizen in case of wrong implementation could be devastating. Introspection may be the safest way to implement a hybrid human-AI system, especially to experiment with and test how well it performs before implementing it completely or scaling it up to a routing level of automation with a higher efficiency gain. When looking at case PP, it can be seen that a higher efficiency gain is wanted as long as there is a minimal level of human control in which the civil servant can intervene with the process. This would result in a routing or even an advisory level of automation. If the technical opportunities allow it, the data that is being used is free of any bias and if the system has been tested in a level of automation of advising or routing this could maybe even be scaled up to a decision-support level of automation as long as the civil servant is in control and, based on the interviews,

Table 5.8: Implementation of Efficiency Gain and Human Control in the Levels of Automation

Level of Automation	Efficiency Gain	Human Control
1. *Introspection*	Efficiency increases slightly for the more complex cases. These cases will be labelled and discussed on why these are more complex and to try and make these decisions go more fluent as well.	The civil servant takes the decision though the AI is running in the background and can be used in cases that are more complex or where more disagreement exists between civil servants. The AI can have an indirect impact on those decisions when the outcome of the AI for those cases is discussed in general meetings.
2. *Routing*	Efficiency increase in the sense that civil servants need less time to look thoroughly at all the criteria since the AI does this as well and identifies the more complex decisions. Therefore the civil servant can go faster through the less complex cases	The civil servant takes the decision while advised by the AI to focus more on certain cases. Here, the AI suggests some more critical cases to the civil servant.
3. *Advisory*	Efficiency increases more than in *routing* since the civil servant gets advice from the AI before taking a decision.	The civil servant takes advice from the AI and looks at the criteria that seem to play the most important role or that are still unsure, before implementing a decision.
4. *Decision-Support*	The efficiency increases since the AI suggest a decision which the civil servant just has to confirm after checking potential critical criteria.	The AI determines the decision whereas the human has the final call and determines whether they implement the decision made by the AI.

is being triggered to critically think on the outcome of the system and of their own reasoning with potential bias in it as well.

One thing needs to be made clear here. Even though table 5.8 mentions four implementations, multiple levels can be thought of or experimented with. For example for the level of automation that is called 'Introspection' in tables 4.3 and 5.8, already two slightly different forms could be thought of where the first one is focused on repetitive introspection and the other on a one-time introspection to improve decision-making and learn from the AI. The goal of this research is to show that by first looking at the trade-offs between the important criteria, the context factors, the technological capabilities, the goal of the automation, etc. a responsible level of automation could be the result. Where to implement the algorithm and how it will cooperate with the civil servant, looking at the frameworks by Cummings and Bruni [2009] and Parasuraman et al. [2000] and at different theories such as the System Safety Theory by Laveson [2012], will further help in responsible implementing a responsible level of automation and may adjust criteria for this implementation, focusing on the civil servant perspective and the role of human control.

For policy-makers in the public domain willing to experiment with the implementation of different levels of automation with a decision support tool such as AI, a canvas can be created that highlights the most important and general aspects to take into account. Filling in this canvas specific for the decision-making task that is to be automated and the complications and opportunities that come with it helps in providing a clear overview of what the goal is, why the automation can be beneficial, who the important stakeholders are, who share the responsibility and how to implement it responsibly. Analysis of the trade-offs that people make in this specific decision-making process can be done through a vignette experiment as conducted in this research, or with another tool. This canvas is briefly explained in the next section and helps in providing guidelines to those willing to experiment with the digitalisation and implementation of AI in the public domain.

5.2.4 Survey Conclusion

In this vignette experiment, the goal is to identify the trade-offs that experts from different domains make regarding the criteria for human control and efficiency gain for less sensitive and more sensitive decision-making cases. The identified trade-offs between efficiency gain and human control have to do with other external context factors or other criteria. For example, for less sensitive cases, a higher efficiency gain can result in more acceptance from the citizens as the waiting time decreases, whereas for more sensitive cases a higher level of human control is preferred where efficiency gain is not considered to play an important role. The main trade-offs are listed below:

- The higher the efficiency gain, the more likely the lower the level of human control will be. As human control takes more time than AI controlled systems.
- The more sensitive a decision, the higher the level of preferred human control
- The higher the impact on the citizen, the more interpretation and thus human control is preferred.
- AI should not result in higher efficiency gain but in better human control as it could be implemented as a mirror where the civil servant can improve the quality of their own decision-making.
- When automation is used for routing more complex decisions, a lower level of human control may be possible. Though AI should not take the decision, but merely select the more complex cases.
- As long as civil servants learn from the AI, any efficiency gain is preferred and thus a lower level of human control is possible as long as the quality increases.
- Due to the high level of sensitivity, a high level of human control with a transparent and controllable system is mandatory

Based on this research, sensitivity is considered too have a big impact on the level of human control and thus on the level of automation that one considers responsible to be implemented. Though for cases where there is less sensitivity the public domain seems to value the efficiency gain higher than the level of automation, at least when this is significant. When the efficiency has barely improved, the question arises of why to even automate at all. So human control is necessary but if there is little efficiency gain, then there might be no need for any automation at all. When this improvement in efficiency is significant depends on the particular case where you want to implement automation.

Next to these trade-offs and context factors, some other remarks can be made as found in this research. Taking the civil servants into the design process will aid in identifying this. Looking at the different domains that filled in the survey, although the number of responses is limited, some interesting remarks were made where the scientific domain, followed by the public domain, was the most enthusiastic and optimistic about automation in a very sensitive decision. Although when considering a less sensitive case, this order had switched and the private domain seemed to have more trust in these systems. Looking at other decision-making processes in other governmental executing organisations the same is expected although more research on the other criteria and context factors is needed in order to come up with a complete picture that can result in a responsible implementation of AI.

5.3 CANVAS

This research consists of desk research, interviews and a vignette experiment in responsibly implementing a level of automation focusing on the civil servant as the meaningful human control. This is all part of coming to a responsible level of automation, but as mentioned in chapter 2, other aspects also play a role. A tool that can be used to determine a responsible level of AI is a canvas. This however needs further research, though this research could be seen as part of that canvas that can be created to support policy-makers willing to experiment with the implementation of AI in the public domain. Other aspects that are included in this canvas are laws and regulations (European guidelines), risk and safety management, stakeholder involvement and questions and where and when in the decision-making process to implement a level of automation. A setup of this canvas can be seen in appendix D. This canvas aims to guide through an overview of aspects of responsibility, technology and a multi-actor perspective, which could contribute to a more responsible, generally accepted and understood level of automation by implementing AI in the decision-making processes of civil servants. The goal of implementing AI should be to provide support to the civil servant and a higher quality of service to the public which may entail an increase in efficiency but foremost should establish a certain level of meaningful human control. This canvas is a first draft and does not aim to provide an answer to the research question, it does however provide a first version of a tool that can be used and tested when having to determine a responsible level of automation in decision-making, specifically for the public domain. This research shows how different stakeholders perceive the trade-offs that arise between criteria when looking at different levels of automation. Working from those criteria and the trade-offs to a level of automation is considered a more responsible way of implementing a level of automation than the other way around where AI is implemented and the criteria are observed and trade-offs are managed. The setup of this canvas can be found in appendix D. The canvas is mainly proposed as a tool for the public domain with a scientific foundation and a testing format that can be applied to AI from the private domain.

5.4 CONCLUSION

As can be concluded from the interviews automation of decision-making in the public domain cannot happen responsibly without the presence of a civil servant who (meaningfully) controls the decision-making process. Following the trade-offs made by the respondents of the vignette experiment, human control is evidently more important than efficiency gain, especially for very sensitive cases. For less sensitive cases there actually seems to be a slight tendency towards a milder form of human control as long as it gives some efficiency gain. As the scenario with more efficiency gain and less human control is slightly preferred by the respondent from the public and private domain. Meaningful human control should contain the aspects of the System Safety Theory as should the civil servant understand the implemented system, have trust in the implemented system and feel responsible for the decision they make together with the implemented system. The policy maker who decides on the implementation of AI or even the Ministry of that domain should be ultimately responsible for any issues in the end, though the civil servant and the creator of the AI also have responsibilities on respectively the proper usage and design of the system.

Although the number of respondents was limited, and thus more testing and experimenting with sensitivity and the trade-off between human control and efficiency gain, some interesting remarks have been made by the respondents and the experts that will build on this qualitative research. Aiming to provide guidance for responsible AI implementation in the public domain, focusing on the aspect of meaningful human control by the civil servant. A final conclusion should be made on the aspect of the impact that automating a decision-making process has on the citizen, as all expert interviewees have mentioned this as either criteria or a context factor that needs to be taken into account. Where the impact overlaps with the sensitivity of the decision, it is worth investigating this more. Respondents from the vignette experiment suggest that more human control can result in more a

6 DISCUSSION AND CONCLUSION

In this thesis, the criteria for determining a responsible level of automation have been examined and the main trade-offs have been identified to provide guidance to those willing to experiment with AI in the public domain. In this chapter, the main findings are discussed and reflected upon, answering the research question. Lastly, a final conclusion is given, with remarks on the limitations of this study and some future study recommendations.

6.1 MAIN FINDINGS RECOMMENDATIONS

Overall, the research sheds light on the nuanced considerations that must be taken into account when implementing systems that rely on both human and algorithmic decision-making. Especially in those decision-making processes where the individual citizen (or groups of citizens) are the direct target of the decision-making. By combining the answers to the sub-research questions, the main research question as repeated below can be answered.

What trade-offs are made between efficiency and human control when determining a responsible level of automation for different decision-making processes in local governments?

This research focuses on the implementation of AI and mainly on the civil servant who has to fulfil the role of human controller when AI is implemented in hybrid human-AI cooperation. Delimited to the decision-making process of smaller, more local government institutions such as municipalities. The aim of this research is to provide guidance for those willing to experiment with responsible human-AI decision-making in the public domain, with a focus on the trade-offs between efficiency gain and human control when implementing AI.

The public domain is known to adapt slower to technological advancements and innovations than the private domain. Especially when considering the implementation of AI tools. Using AI tools in the public domain brings plenty of challenges as this domain is not merely held accountable based on their profit, and maybe their quality of service but also on their responsiveness, their equality of service, their fairness and openness on any decision made. Implementing AI in the decision-making process by civil servants is, therefore, a very tricky thing to do as AI has been famous for its black box principle. This research aims to provide a bottom-up approach by looking at a specific decision-making task and identifying the criteria that play a role in determining a responsible level of automation for this specific task. The desk research and interviews did give a more complete overview of the other aspects that need to be taken into account. Therefore a setup for a canvas has been created at the end of this research to provide insights into how this analysis of criteria and context factors can aid in achieving a responsible level of automation with human-AI decision-making in the public domain. Already focusing on future study directions.

First desk research has been conducted to investigate all the aspects that need to be taken into account when thinking about implementing AI in public decision-making. The most important aspects were the identification of criteria that differ per level of automation and that could be used to determine a responsible and appropriate level of automation. These criteria and their possible trade-offs were analysed through interviews and a vignette experiment, as being a part of the entire policy-making process when deciding upon a responsible and appropriate level of automation. The canvas provided at the end of the results chapter gives some insights into the other aspects that were not extensively discussed and researched, but do provide additional viewpoints, theories and frameworks to consider when gathering all the information in order to make a well and thoroughly considered decision. Three domains have been the target of this investigation as these three have the main expertise and experience when discussing the responsible implementation of AI in the public domain. Recommendations can be made on a number of levels as presented in table 6.1 at the end of this chapter.

6.1.1 Hybrid Human-AI Systems

From the literature a number of aspects arose that should be taken into account when looking at the correct level of automation. As already extensively discussed, implementing a responsible level of automation needs a responsible level of human control, where we, as humans, are still in control over the decision. AI has the benefit to make processes more efficient but we also need to improve quality, especially in the public domain. AI can be used to show our own biases and to learn from these biases. It should be used in a manner, especially for more complex and sensitive decision-making, where it enables civil servants to become a better version of themselves. Replacing civil servants is not yet realistic. As resulted from the interviews, automating decision-making or parts of these processes should improve the individual decision-making by civil servants resulting in better public service, and it should provide the civil servant with a more diverse and well-balanced set of work tasks where more emphasis has to be placed on direct communication and care with citizens. Civil servants want to help the citizens, and they want to provide the best care possible, which requires personal attention. This would be an ideal situation, increasing efficiency and quality by implementing AI as a mirror for the civil servant and giving the civil servant a certain level of human control to ensure quality.

6.2 TRADE-OFFS

From chapter 4 can be concluded that the three criteria efficiency gain, human control and quality were the most prominent ones mentioned in the literature but also most often mentioned in the interviews, as validated in chapter 5, section 5.1. The main trade-off was found between efficiency gain and human control. Human control was however mentioned by all the experts in the interview as being the vital criterion. A high level of automation obviously has a low level of human control as the decision is taken with more influence of the AI and less of the civil servant. All hybrid levels of automation have a certain form of control from the human and the AI.

The experts and respondents stated that there should always be a certain level of human control. This, however, goes at the cost of efficiency gain since (for now) no full automation can be fully controlled by a human being simply because we do not fully understand how most AI systems come to a decision or how it exactly analyses the data, making it hard to provide an explanation to a citizen which is obligated by law. For hybrid human-AI systems, the efficiency gain can be lowered for a certain level of automation when they prefer more human control. For example in the level of automation of a more advisory role, more human control could be

Table 6.1: Recommendations for Main Research Blocks

Research aspect	Recommendation	Based on Main Finding
Criteria	• Efficiency gain should for sensitive cases not be prioritised over human control. • Any decision-making task should have a minimum level of human control. • Human control should be the most important thing to consider when deciding upon a level of automation, where AI should be implemented to support and improve the decision-making of the civil servant. Implement AI as a *"mirror"* for civil servants.	• Respondents of the survey are hesitant over scenarios where the efficiency is high but the level of human control is not really high or slightly vague. • During the interviews and the survey it became clear that as soon as there is a sufficient level of human control where the civil servant has a good understanding of what happens within the AI and can intervene with it, AI implementation is considered responsible. This would automatically exclude full automation and the higher levels of hybrid automation. • The interviewees stated that a clear goal for the implementation of AI needs to be present where a focus should be on the support and improvement of the decision-making by the civil servant. This could also be
Context factors	• The more sensitive a decision the more human control should be present when deciding upon a level of automation. • There should always be a possibility for the civil servant to intervene, even if the decision-making process is easy to automate and not that sensitive. • The impact of a decision that is made by a hybrid human-AI system on the individual citizen should be minimised and well-understood.	• From the survey, it became clear that for sensitive decisions, with possibly a high impact on the individual citizen, a high level of human control is wanted even if the efficiency has increased a lot. • For less sensitive cases, the efficiency gain and human control are both considered interesting criteria, though there is not a sufficient level of human control in the sense that the civil servant should always have the ability to intervene with a decision and question the outcome of the algorithm. • The impact is mentioned in all interviews as a factor that should determine the level of automation. This overlaps with aspects of vulnerability, sensitivity, human control, system safety, etc. and should be well addressed through for example scenario analysis, impact assessment and risk analysis.
Responsibility	• There should be a shared responsibility for the decision-making when a level of automation has been implemented, shared between the policy maker, the algorithm developer and the civil servant. • The civil servant is responsible for a safe decision that is being made. • The AI can never be responsible for a wrong or damaging decision. • the board of directors or the minister is ultimately responsible for the proper, fair and equal function of a hybrid human-AI system when implemented.	• The policymaker decides on whether to implement AI in any way. The operator provides an AI that should be suitable for the implementation in decision-making in the public domain and should not be biased, black box, etc. The civil servant has to interpret the outcome of the decision made by the AI and is thus responsible for a fair and just decision. Being critical of the algorithm and of themselves. • Because the civil servant should always have the possibility to intervene, and they are expected to always question the outcome of the AI. Thus they can be considered to be responsible for the final decision. If there is a malfunction in the way the AI is implemented or if the AI is giving biased advice, the civil servant would have done their best in detecting this. • The operator of the AI may be responsible to an extent for the function of the AI, especially when it doesn't function as agreed on or when it does appear to have biases that it was supposed to be free of. Communication and understanding of the algorithm are key here.

preferred where the civil servants have more room for their own interpretation and intervention and have more learning abilities from the algorithm. This may result in a slower system than when relying more on the AI from the start, though over time may be more valuable. Again depending on the goal and the aim of implementing

AI. The criteria listed below resulted from the interviews and have been tested in the interviews.

The list below sums up the trade-offs as they were identified by experts already in the interviews and as they resulted from the vignette experiment where the respondents were given the room to identify and elaborate on the trade-offs they saw and made.

- When more efficiency gain is preferred, a higher level of automation is more suitable, where this will result in a lower level of human control from the civil servants.
- A higher sensitivity of the decision that is to be made, results in a preferred higher level of human control over efficiency gain. Where the latter is considered less significant.
- When the impact of the decision is higher on an individual level, more human control is preferred over efficiency gain with clear explainability on the decision-making process.
- When implementing the AI on identification of more complex cases instead of automating the decision a higher level of efficiency gain is preferred over human control. Where experts and respondents mentioned the possibility to improve the decision-making of civil servants.
- An increase in either efficiency gain or human control results in a higher acceptance from the public. Context factors determine which criteria should be considered more important.
- More efficiency gain is preferred over human control over time where the AI has stimulated the civil servant to learn from and be more critical of their own decision-making. Thus, decreasing human control may be preferred over time.

Resulting from the interviews and mainly from the vignette experiment, more in-depth and detailed reasons behind the trade-offs between efficiency gain and human control have been found when determining a certain level of automation. Here it can be concluded that a more sensible decision results in a preference for a higher level of human control and less focus on any efficiency gain. In the less sensitive case, the efficiency gain was considered more important though the civil servant needs to have the possibility to intervene, so a form of human control had to be in place according to the respondents. Though when looking at the two scenarios for the non-sensitive case (Parking Permit), there is a big difference in the control that the civil servant has over the decision-making. The third scenario allows the civil servant to accept or not accept a decision made by the AI, whereas in scenario four the civil servant has the possibility to adjust the decision based on their findings and conclusions. The latter stimulates the civil servant a lot more in thinking for themselves and considering the context in which the decision has to be made. However, scenario 3 is rated higher in willingness to use than scenario 4. This could be explained by the trust in the capabilities of an AI in more simple and less sensitive tasks. The impact on the individual citizen is suggested to be less of a concern here.

Giving an answer to the second sub-question, regarding efficiency gain and human control it is clear that for sensitive cases a higher level of human control is preferred than for non-sensitive cases. As soon as there is a certain degree of responsible human control, where the civil servant understands the AI and its outcome and can interfere with it, any efficiency gain is considered good and acceptable. However, for the less sensitive cases, a lower degree of human control is wanted with more efficiency gain as these cases are considered to be more easily automated and are more recurrent. With which the respondents point to the fact you don't want to doubt every decision made by the AI, this will only take more time.

For the less sensitive cases, the public domain is a bit more sceptical due to the inconvenience for the civil servant to interfere in these more recurrent decisions. Recurrence was not mentioned as one of the context factors, so more research needs to be done on this to confirm this finding. The private domain is more positive towards the less sensitive cases as they may see less risk for "*their*" AI to fail.

After making these trade-offs between criteria and the impact of context factors into account, one should be able to already scope down the different levels of automation to a subset of levels or maybe already to one level. It must be pointed out that a flexible approach towards these levels is advised. There may be hybrid human-AI cooperation that is in between the levels mentioned in this report or levels that may overlap on two sides. As mentioned earlier, for the level of automation that is mentioned as 'Introspection' one can already think of two slightly different forms where the first one is about repetitive introspection and the other a one-time introspection to improve decision-making and learn from the AI. The idea is that based on the important criteria, their trade-offs and the context factors, a good step can be taken towards a responsible level of automation. This starts to answer sub-research question 3 as presented below.

When looking at the first case of sensitive decision-making, the scenario with a high level of human control and some efficiency gain was preferred over the scenario with a lot of efficiency gain, but very little human control. This would suggest that a lower level of automation would be more responsible and more accepted from a civil servant perspective. As they will still have control over a very sensitive decision and (as mentioned in the interviews) still have the feeling that they can take a good decision for an individual, which is what they want. Civil servants want to help and to aid individual citizens, it's not a production line.

6.3 MEANINGFUL HUMAN CONTROL

This reason mentioned above could be suggested as a reason why meaningful human control is such a topic in this responsible AI puzzle. Civil servants are the human controller, and you want them to feel responsible, to understand the entire process, to understand the human-AI interaction, to understand the outcome, to question it and to learn from it. In order to achieve responsible human control by the civil servant the following aspects need to be considered and/or implemented. Looking at it from a system safety theory perspective, 1) the civil servant needs a clear goal, 2) should be able to deliver input to the system, 3) should know and understand the system, and 4) should be able to observe the system properly. Adding to this theory and based on the results from this research a responsible civil servant 5) should understand their role in the hybrid-AI decision-making system and 6) should be aware of their responsibility in the decision-making. Using the AI as a "mirror", as was mentioned in interview B.8, will also contribute to a more responsible level of human control as the civil servant 7) will be more aware of their own flaws and biases through the use of AI. A final aspect of responsible human control of civil servants in hybrid human-AI decision-making is the task of civil servants to 8) deliver good care and fair, just and equal service to individual citizens. As already mentioned this last requirement for meaningful human control is already in place as the civil servants, according to the expert interviews, always want to provide the best care possible for the citizens.

The policymaker who decides upon the implementation of AI in the decision-making with civil servants plays another big role in reaching meaningful human control. They should involve civil servants in every step of this decision to implement AI and in what way it can best support and improve the work of civil servants. Next to this, there should be a clear understanding of all values that come into play when regarding, such as transparency, equality, explainability, etc. This needs to be agreed on beforehand so that everybody understands the rules of the game. If deviation or different interpretations occur this gives room for debate on improving the definitions of those values. Policymakers should be fully aware of what AI could do, what it involves and how this may impact the decision-making of civil servants.

6.4 ACADEMIC REFLECTION

Before discussing the scientific contributions and societal relevance, it is important to highlight some limitations of this study. Quite some assumptions have been made together with some scoping down in order to really delve into the aspect of human control by civil servants and the trade-off with efficiency gain in different contexts. Though the findings of this research provide valuable insights, the lack of significance demands further research into this topic as will be discussed later.

6.4.1 Limitations

A first note needs to be made on all the results from this research as no statistical significance has been found for any of the results from the vignette experiment. Though most findings could be explained or built on using expert interviews. It is important to be aware of this lack of significance and conduct further research into the responsible implementation of AI and the trade-offs between criteria. Needless to say, the number of respondents was one of the main reasons for a lack of significance and thus this is something that needs to be taken into account carefully when analysing or using the results from this research.

As a second limitation, and as mentioned in the results section, it is important to clearly define what is meant by values such as equality, good implementation, fairness, transparency, etc. in order to be able to reach those values. Though the definition of those values can differ for example when looking at public organisations in other countries for example Asia or Africa. Some values might not even be considered important in other parts of the world, as can be the same for the identified criteria and trade-offs. Generalising the results from this research will probably extend not far beyond the Dutch borders. But should be foremost focused on implementing responsible levels of automation for different decision-making tasks within the Dutch government.

Thirdly, the experts asked for an interview and the respondent of the survey all had prior knowledge about the topic and feel engaged with the topic of responsible AI implementation in the public domain. Asking them to view the selection of a responsible level of automation based on a trade-off between the two main criteria from a civil servant perspective, may give a slightly distorted picture as this may not be in line with what the public or even with what the civil servants think. They were also collected through snowballing and personal communications which could have resulted in a group of like-minded experts which would even further have increased the bias in the data. Although the experts do show some differences in perspectives and especially in their substantiation, more research is needed with more diverse and different groups of experts and respondents. As for the respondents they were all approached at the same congress which could also have resulted in a one-sided group with bias as a result.

Furthermore, the findings from the survey are hard to generalise as the group of respondents was very small. Also because of this, no randomisation had been implemented which results in less quantitative results. Also, it focuses more on generalising the results to other public decision-making processes than generalising for the population. These qualitative findings should be tested more and maybe in different aspects of the public domain and in different public organisations, though they already do provide some interesting insights which contribute to a first step in creating an overview or a framework that could be used for those willing to experiment with AI implementation in the public domain.

Next to this, when looking at the HACT model, as presented in Chapter 4, by Cummings and Bruni [2009], this research focuses on the middle section of the model where algorithms or AI can be implemented in automating the decision-making process. There is however no focus on using automation tools in the data acquisition or the executing sections. The latter could be excluded due to the level of human control that is preferred to be present but the data acquisition could be suggested as an extra addition to this research where probably a lot of efficiency gain can be achieved. Finally, a notion has to be made on some value judgement that may have been made in the survey that was not made by the respondents but rather for them. As some subjective words may be used that can be differently interpreted by the respondents as "clear" and "sensitive data". With the latter, the respondents were presented that a parking permit uses less sensitive data than special welfare, which may not be the case for certain respondents as data on your address, etc may be considered equally important as data on your financial situation.

As a final remark on the limitations, it can not be stressed enough that the conclusions and recommendations drawn from this research should be analysed and used carefully and with awareness of the lack of significance. Although they provide valuable insights on the trade-offs between efficiency gain and human control, adding to guidelines that could be used for choosing a responsible level of automation, further research is absolutely needed to confirm these findings and to be able to give advice on a responsible level of automation.

6.4.2 Scientific Contribution

More and more is being written every day on the (responsible) implementation of AI and (meaningful) human control. This research uses interviews and a vignette experiment to give qualitative results on what experts think should be important to consider when looking at human control from a civil servant perspective. This research contributes to a more concrete advice and manner to determine a responsible level of automation in the public domain, focusing on the side of the civil servant, that is the operator and thus the human control in hybrid human-AI decision-making systems. Including the civil servant in the design of this decision-making, system is key to reaching a responsible and meaningful level of human control. This will also aid in creating more acceptance from the civil servants but also from the public, building trust in these decision-making processes supported by algorithms.

This research also tries to build on the concept of meaningful human control as often mentioned in the debate on implementing human-AI systems. By giving 8 requirements that the civil servant should adhere to and should be consciously aware of, the concept of meaningful human control for the civil servant becomes more concrete. Responsible implementation of AI has been and will remain a topic of interest since the developments in this new technology are going faster than our regulations and ways of working with AI. This research aims to provide insights into multiple aspects that need to be considered for a responsible implementation of AI specifically in the public domain. Though with a main focus on the criteria efficiency gain and human control, that determine a responsible level of automation.

6.4.3 Societal Relevance

Responsible implementation of AI, specifically in the public domain, has the potential to create better and more equal public service with an increase in quality and efficiency when it is implemented in a way where the civil servant can learn from it and uses it as a mirror. Through this, civil servants become aware of their own biases and of their own misjudgements. This could benefit when both AI and civil servants complement each other with their own skills. When implementing AI in the decision-making processes results in more time for the civil servant, they can spend this on more one-on-one communication and interaction with citizens, resulting in more human-centred public service. Though it is key here that there is a shared responsibility, as suggested by the experts, where the civil servants feel responsible for the outcome of the decision, the developer of the AI for how the AI functions, and the policy maker on the entire decision-making. The latter is responsible in case something goes wrong or when there is a malfunction in the hybrid human-AI decision-making process. These recommendations on the aspect of responsibility can be seen in table 6.1.

6.5 RELEVANCE TO STUDY

According to the previous Director of Studies of the Master programme Engineering & Policy Analysis (EPA), Bert Enserink, the topic of an EPA MSc thesis should be linked to one or multiple of the Grand Challenges as discussed in the study programme [Enserink, 2017]. The thesis is supposed to focus on situations where policy is lacking or underdeveloped, and where more information is needed to contribute to society and to get a better implementation of the subject at hand. Adequate policy-making is failing, since the public domain is still lacking behind in the implementation of AI. This master thesis of EPA is related to the grand challenge of digitalisation overlapping with aspects of safety and security challenges regarding privacy and responsibility issues when using new information technologies in the public domain. Especially the aspect of the civil servant perspective on the implementation of different levels of automation and their trade-off, not enough research has been conducted so far, let alone the lack of guidance from a European or national level.

6.6 FUTURE RESEARCH SUGGESTIONS

Due to the small number of respondents that were asked to fill in the vignette experiment, there is definitely a need for more research on the trade-offs between the other criteria for determining a level of automation as well. Especially when looking at different decision-making processes in different scenarios with other context factors these trade-offs can be very interesting. It is therefore also advised to use the canvas as a tool for determining a responsible level of automation for a certain decision-making process in the public domain. Depending on the specific task that the AI will be used in, these different aspects should be approached differently. This will result in a tailored level of automation. This research has focused on one part of the canvas which is considered the most important and where the least research has been conducted in. Quantifying values or these criteria can be quite hard especially when looking at their trade-offs. In this research, a vignette experiment has been used to look at the trade-offs of a very small set of criteria with only one context factor. Including more would result in larger sets of vignettes and other research methods could be used here. It is advised that this canvas is further developed and also tested in workshops to gain feedback from experts in the three domains but also from civil servants and citizens to understand how the decision for a certain level of automation is made and what considerations will be made.

As can also be concluded by the results chapter, the impact that the automation of decision-making by civil servants has on individual citizens is a very important aspect of the problem that needs to be taken into account. This overlaps slightly with the contextual factor of sensitivity as discussed in this research, but more research is necessary also on the citizens' perspective. Impact assessments together with risk assessments (as mentioned in the canvas) need to be made in order to determine how badly individuals can be impacted in case of the wrong implementation of AI. This would also help to provide insights to individual citizens, to better test their perspective on the automation of public decision-making tasks, and what this entails for them.

Adding to this, conducting more empirical research to investigate the balance between human control and efficiency gain could bolster trust in the empirical accuracy of the theory, while also yielding a more nuanced comprehension of the impact of this trade-off. Also when incorporating the perspective of the civil servants, it should be mentioned that there is a need to include the civil servants in reaching a responsible AI implementation. For this research these have not been included though the experts and respondents for the vignette experiment (of whom some were civil servants) were asked to give their ideas from a civil servant perspective, focusing on human control. Not to forget, including European and national laws and regulations in future research will result in a more complete picture with hopefully clear, refreshing and helpful insights for policymakers.

Finally, a set-up for a canvas has been presented as a way of identifying a responsible level of automation by taking into account all the relevant aspects of this problem. This research has aimed at one "*piece-of-the-puzzle*" but further research is needed in the other aspects as mentioned and maybe even more or different ones. This canvas has the aim to provide a way of going about this research and is merely presented as a tool for including the decision-makers (and preferably also the civil servants) in the implementation of a good and responsible policy. Though other tools could be used to analyse aspects of this problem and to identify a responsible level of automation to support civil servants.

6.7 concluding remarks

Responsible implementation of AI in the public domain can only be established through good communication and clear rules where all involved stakeholders have a mutual understanding of these rules. This includes the clear definition of values and requirements, where debates can be held on the ambiguity of these concepts in order to gain a mutual understanding. Looking at the responsibility aspect, this also includes taking all stakeholders along in the policy implementation process where a level of automation is being determined. Dividing the responsibility among the three actors as provided in this research can help in building on the acceptance by both the citizens and the civil servants, though it must be clear that the one who decides to implement a form or level of automation should be the one ultimately responsible for any malfunction. Regarding the usage of AI at any of the levels as suggested in this report, it is advised to look at AI as a supportive tool to help civil servants in increasing their quality of decision-making and improve efficiency with meaningful human control by the civil servants. Presenting AI as a sort of mirror for civil servants enables them to learn from the AI and to be more aware of their own biases and mistakes. This will eventually help in increasing efficiency and quality but also improve the social interaction that civil servants have with citizens as they can better explain what went wrong, or how a decision is made. A certain level of transparency and explainability of the AI is however required to achieve this.

Before implementing a responsible level of automation it is advised to look at an experimenting or pilot phase where the level of automation can be tested, also for the civil servant to get adjusted to it. Running the AI next to the normal decision-making can also already be seen as the ”mirror” as mentioned above. This may in the beginning not result in more efficient decision-making over time, but the quality will rise and as the civil servants become more familiar with AI and they develop more trust in using these tools, efficiency can be increased. There should be a fine balance in which the citizen eventually benefits from the best of both worlds regarding civil servants (human) and AI in the hybrid human-AI decision-making, complementing each other. More efficient decision-making through the implementation of AI improves the handling of tickets at governmental institutions but with a level of meaningful human control where the civil servant can have more time for direct one-on-one contact with the citizen providing explanations, reasoning and help. Which should eventually be the role of the civil servant.

BIBLIOGRAPHY

Agostino, D., Saliterer, I., and Steccolini, I. (2022). Digitalization, accounting and accountability: A literature review and reflections on future research in public services. *Financial Accountability and Management*, 38:152–176.

Aguinis, H. and Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodology studies. *Organizational Research Methods*, 17:351–371.

Algoritmeregister (2020). Het algoritmeregister van de nederlandse overheid.

Andronie, M., Lăzăroiu, G., Iatagan, M., Uță, C., Ștefănescu, R., and Cocoșatu, M. (2021). Artificial intelligence-based decision-making algorithms, internet of things sensing networks, and deep learning-assisted smart process management in cyber-physical production systems. *Electronics (Switzerland)*, 10.

Araujo, T., Helberger, N., Kruikemeier, S., and de Vreese, C. H. (2020). In ai we trust? perceptions about automated decision-making by artificial intelligence. *AI and Society*, 35:611–623.

Arnaboldi, M., de Bruijn, H., Steccolini, I., and der Voort, H. V. (2022). On humans, algorithms and data. *Qualitative Research in Accounting and Management*, 19:241–254.

Asadi, E., Silva, M. G. D., Antunes, C. H., Dias, L., and Glicksman, L. (2014). Multi-objective optimization for building retrofit: A model using genetic algorithm and artificial neural network and an application. *Energy and Buildings*, 81:444–456.

Atzmüller, C. and Steiner, P. M. (2010). Experimental vignette studies n survey research. *Methodology*, 6:128–138.

Bannister, F. and Connolly, R. (2020). Administration by algorithm: A risk management framework. *Information Polity*, 25:471–490.

Bharosa, N. (2022). The rise of govtech: Trojan horse or blessing in disguise? a research agenda. *Government Information Quarterly*, 39.

Bolderston, A. (2012). Conducting a research interview. *Journal of Medical Imaging and Radiation Sciences*, 43(1).

Bullock, J. B. (2019). Artificial intelligence, discretion, and bureaucracy. *American Review of Public Administration*, 49:751–761.

Callamard, A. (2023). Responsible artificial intelligence in the military domain. Opening talk of the conference.

Canhoto, A. I. and Clear, F. (2020). Artificial intelligence and machine learning as business tools: A framework for diagnosing value destruction potential. *Business Horizons*, 63:183–193.

Carlizzi, D. N. and Quattrone, A. (2023). The algorithmic public decision, between explainability, administrative discretion and data-driven decision making. *Artificial Intelligence and Economics: the Key to the Future*, 523:123–135.

Carter, L. and Bélanger, F. (2005). The utilization of e-government services: Citizen trust, innovation and acceptance factors. *Information Systems Journal*, 15:5–25.

Cath, C. and Jansen, F. (2023). Dutch comfort: The limits of ai governance through municipal registers.

Cockburn, I. M., Henderson, R., and Stern, S. (2019). *The Impact of Artificial Intelligence on Innovation : an Exploratory Analysis*.

Combes, P.-P., Duranton, G., Gobillon, L., Puga, D., and Roux, S. (2012). The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica*, 80:2543–2594.

Coombs, C., Hislop, D., Taneva, S. K., and Barnard, S. (2020). The strategic impacts of intelligent automation for knowledge and service work: An interdisciplinary review. *The Journal of Strategic Information Systems*, 29:101600.

Crespo, A. M. F. (2019). Less automation and full autonomy in aviation, dilemma or conundrum? *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 2019-October:4245–4250.

Cummings, M. L. and Bruni, S. (2009). Collaborative human–automation decision making. *Springer Handbook of Automation*.

de Boer, N. and Raaphorst, N. (2021). Automation and discretion: explaining the effect of automation on how street-level bureaucrats enforce. *Public Management Review*.

de Mello, L. and Ter-Minassian, T. (2020). Digitalisation challenges and opportunities for subnational governments luiz de mello and teresa ter-minassian oecd working papers on fiscal federalism.

de Sio, F. S. and van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers Robotics AI*, 5.

Delfos, J. (2022). The accuracy-explainability trade-off of machine learning: A stated preference experiment for the case of border control.

Delfos, J., Zuiderwijk, A., Cranenburgh, S. V., and Chorus, C. (2022). Perceived challenges and opportunities of machine learning applications in governmental organisations: an interview-based exploration in the netherlands. *ACM International Conference Proceeding Series*, pages 82–89.

Dey, S. and Lee, S.-W. (2021). Multilayered review of safety approaches for machine learning-based systems in the days of ai. *Journal of Systems and Software*, 176:110941.

Dhungel, A. K., Wessel, D., Zoubir, M., and Heine, M. (2021). Too bureaucratic to flexibly learn about ai? the human-centered development of a mooc on artificial intelligence in and for public administration. *ACM International Conference Proceeding Series*, pages 563–567.

Dieter O., K. (2022). Netherlands ranked as eu's 3rd most digitalised country. *I Amsterdam Business*.

Dobbe, R. I. J. (2022). System safety and artificial intelligence.

EC (2019). A definition of ai: Main capabilities and disciplines.

EC (2021a). Proposal for a regulation of the european parliament and of the council: Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts.

EC (2021b). Regulation of the european parliament and of the council.

EC (2022). Regulatory framework proposal on artificial intelligence.

Elliott, K., Price, R., Shaw, P., Spiliotopoulos, T., Ng, M., Coopamootoo, K., and Moorsel, A. V. (2021). Towards an equitable digital society: Artificial intelligence (ai) and corporate digital responsibility (cdr).

Endsley[2], M. R. and Kaber, D. B. (1999). Level of automation eoe ects on performance, situation awareness and workload in a dynamic control task.

Engin, Z. and Treleaven, P. (2019). Algorithmic government: Automating public services and supporting civil servants in using data science technologies. *Computer Journal*, 62:448–460.

Enserink, B. (2017). Epa thesis requirements.

EPRS (2022). Ethical and societal challenges of the approaching technical storm.

Evans, S. C., Roberts, M. C., Keeley, J. W., Blossom, J. B., Amaro, C. M., Garcia, A. M., Stough, C. O., Canter, K. S., Robles, R., and Reed, G. M. (2015). Vignette methodologies for studying clinicians' decision-making: Validity, utility, and application in icd-11 field studies. *International Journal of Clinical and Health Psychology*, 15:160–170.

Februari, M. (2023). Democratie is geen product, maar een gemeenschappelijk proces. en dat wankelt door ai. *NRC opinie*.

Feuerriegel, S., Dolata, M., and Schwabe, G. (2020). Fair ai. *Business & Information Systems Engineering*, 62:379–384.

Fiebag, J. (2022). Analysing and predicting the parking occupancy of micromobility vehicles on sidewalks. *openresearch.Amsterdam*.

Gesk, T. S. and Leyer, M. (2022). Artificial intelligence in public services: When and why citizens accept its usage. *Government Information Quarterly*, 39.

Goodman, L. A. (1961). Snowball sampling. *The Annals of Mathematical Statistics*, 1:148–170.

Goodrich, M. and Boer, E. (2000). Designing human-centered automation: trade-offs in collision avoidance system design. *IEEE Transactions on Intelligent Transportation Systems*, 1:40–54.

Gornishka, I. and Sukel, M. (2022). Smart mobility and crowdedness management. *Redactie Openresearch.Amsterdam*.

Gotthardt, M., Koivulaakso, D., Paksoy, O., Saramo, C., Martikainen, M., and Lehner, O. (2020). Current state and challenges in the implementation of smart robotic process automation in accounting and auditing. *ACRN Journal of Finance and Risk Perspectives*, 9:90–102.

Gulenko, A., Kao, O., and Schmidt, F. (2020). Anomaly detection and levels of automation for ai-supported system administration. *Information Management and Big Data*, pages 1–7.

Hadwick, D. and Lan, S. (2021). Lessons to be learned from the dutch childcare allowance scandal: A comparative review of algorithmic governance by tax administrations in the netherlands, france and germany. *World Tax Journal - Amsterdam*, 13:609–645.

Hedström, K., Kolkowska, E., Karlsson, F., and Allen, J. P. (2011). Value conflicts for information security management. *Journal of Strategic Information Systems*, 20:373–384.

Heverly, M. A., Fitt, D. X., and Newman, F. L. (1984). Constructing case vignettes for evaluating clinical judgment: An empirical model. *Evaluation and Program Planning*, 7:45–55.

Hood, C. (1991). Public administration vol. 69:3–19.

Ivanov, S. H. (2022). Automated decision-making. *Foresight*.

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., Riemsdijk, M. B. V., and Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, 3:43.

Karimian, G., Petelos, E., and Evers, S. M. A. A. (2022). The ethical issues of the application of artificial intelligence in healthcare: a systematic scoping review. *AI and Ethics*.

Khan, Z., Ali, M., Kirikkaleli, D., Wahab, S., and Jiao, Z. (2020). The impact of technological innovation and public-private partnership investment on sustainable environment in china: Consumption-based carbon emissions analysis. *Sustainable Development*, 28:1317–1330.

Kokkeler, B. (2022). Vng principes voor de digitale samenleving 2022. *Advies Ethische Commissie - 'ongevraagd'*.

Kool, L., Poel, M., and Giessen, A. V. D. (2010). How to decide on the scope, priorities and coordination of information society policy? analytical framework and three case studies.

Kuziemski, M., Mergel, I., Ulrich, P., and Martinez, A. (2022). Govtech practices in the eu.

Kuziemski, M. and Misuraca, G. (2020a). Ai governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy*, 44.

Kuziemski, M. and Misuraca, G. (2020b). Ai governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings. *Telecommunications Policy*, 44.

König, P. D. (2022). Citizen conceptions of democracy and support for artificial intelligence in government and politics. *European Journal of Political Research*.

König, P. D. and Wenzelburger, G. (2021). The legitimacy gap of algorithmic decision-making in the public sector: Why it arises and how to address it. *Technology in Society*, 67.

Langer, P. C. (2016). The research vignette. *Qualitative Inquiry*, 22:735–744.

Laveson, N. G. (2012). *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press.

Leeuw, A. (2022). Akkoord over europese wet die ai reguleert. *Binnenlands Bestuur*.

Lindgren, I., Østergaard Madsen, C., Hofmann, S., and Melin, U. (2019). Close encounters of the digital kind: A research agenda for the digitalization of public services. *Government Information Quarterly*, 36:427–436.

Lockey, S., Gillespie, N., Holm, D., and Someh, I. A. (2021). A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions.

Luk, S. C. Y. (2009). The impact of leadership and stakeholders on the success/-failure of e-government service: Using the case study of e-stamping service in hong kong. *Government Information Quarterly*, 26:594–604.

Makoza, F. (2019). Developing a taxonomy for identifying stakeholders in national ict policy implementation. *International Journal of RD Innovation Strategy*, 1:44–65.

Maragno, G., Tangi, L., Gastaldi, L., and Benedetti, M. (2022). Ai as an organizational agent to nurture: effectively introducing chatbots in public entities. *Public Management Review*, pages 1–31.

Marjanovic, O. and Cecez-Kecmanovic, D. (2017). Exploring the tension between transparency and datification effects of open government is through the lens of complex adaptive systems. *Journal of Strategic Information Systems*, 26:210–232.

Martinho, A., Kroesen, M., and Chorus, C. (2021). A healthy debate: Exploring the views of medical doctors on the ethics of artificial intelligence. *Artificial Intelligence in Medicine*, 121.

McCarthy, J. (2007). From here to human-level ai. *Artificial Intelligence*, 171:1174–1182.

McLellan, E., MaCqueen, K. M., and Neidig, J. L. (2003). Beyond the qualitative interview: Data preparation and transcription. *Field Methods*, 15:63–84.

Mehmood, B. S., Naseer, S., Chen, D., Seror, A., Sigstad, H., Goldfarb, A., Karo, A., Hjort, J., Kim, W., Vivalt, E., Reenen, J. V., Vautrey, P.-L., Tariq, S., Khalid, S., and Ali, B. (2023). The ripple effect of ai training on policymakers and citizens: Unintended consequences in a developing nation.

Mehr, H. (2017). Artificial intelligence for citizen services and government.

Meinert, E., Alturkistani, A., Foley, K. A., Osama, T., Car, J., Majeed, A., Velthoven, M. V., Wells, G., and Brindley, D. (2019). Blockchain implementation in health care: Protocol for a systematic review. *JMIR Research Protocols*, 8.

Mendoza, J. M. F., Gallego-Schmid, A., Velenturf, A. P., Jensen, P. D., and Ibarra, D. (2022). Circular economy business models and technology management strategies in the wind industry: Sustainability potential, industrial challenges and opportunities. *Renewable and Sustainable Energy Reviews*, 163:112523.

Methnani, L., Tubella, A. A., Dignum, V., and Theodorou, A. (2021). Let me take over: Variable autonomy for meaningful human control. *Frontiers in Artificial Intelligence*, 4.

Mikalef, P., Fjørtoft, S. O., and Torvatn, H. Y. (2019). Artificial intelligence in the public sector: A study of challenges and opportunities for norwegian municipalities. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11701 LNCS:267–277.

Mikalef, P., Lemmer, K., Schaefer, C., Ylinen, M., Fjørtoft, S. O., Torvatn, H. Y., Gupta, M., and Niehaves, B. (2022). Enabling ai capabilities in government agencies: A study of determinants for european municipalities. *Government Information Quarterly*, 39.

Mitrou, L., Janssen, M., and Loukis, E. (2021). Human control and discretion in ai-driven decision-making in government. *ACM International Conference Proceeding Series*, pages 10–16.

Muhlenbach, F. (2020). A methodology for ethics-by-design ai systems: Dealing with human value conflicts.

Nau, D. S. (2009). Artificial intelligence and automation. *Springer Handbook of Automation*, pages 249–266.

Nemitz, P. (2018). Constitutional democracy and technology in the age of artificial intelligence. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376.

Niet, I. A., Dekker, R., and van Est, R. (2022). Seeking public values of digital energy platforms. *Science Technology and Human Values*, 47:380–403.

Nof, S. Y. (2009). Automation: What it means to us around the world. *Springer Handbook of Automation*, pages 13–52.

NordicHQ (2023). The netherlands compared to other european countries.

Ojamo, J. (2021). Use of artificial intelligence by the police: Meps oppose mass surveillance. *News European Parliament*.

Oviedo-Trespalacios, O., Peden, A. E., Cole-Hunter, T., Costantini, A., Haghani, M., Rod, J. E., Kelly, S., Torkamaan, H., Tariq, A., David, J., Newton, A., Gallagher, T., Steinert, S., Filtness, A. J., and Reniers, G. (2023). The risks of using chatgpt to obtain common safety-related.

Parasuraman, R. (2000). Designing automation for human use: empirical studies and quantitative models. *Ergonomics*, 43:931–951.

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans.*, 30:286–297.

Qin, C., Yao, D., Shi, Y., and Song, Z. (2018). Computer-aided detection in chest radiography based on artificial intelligence: A survey. *BioMedical Engineering Online*, 17.

Raaijmakers, S. (2019). Artificial intelligence for law enforcement: Challenges and opportunities. *IEEE Security & Privacy*, 17(5):74–77.

Rafner, J., Dellermann, D., Hjorth, A., Verasztó, D., Kampf, C., Mackay, W., and Sherson, J. (2021). Deskilling, upskilling, and reskilling: a case for hybrid intelligence. *Morals & Machines*, 1:24–39.

Ranerup, A. and Henriksen, H. Z. (2019). Value positions viewed through the lens of automated decision-making: The case of social services. *Government Information Quarterly*, 36:101377.

Reed, M. S., Graves, A., Dandy, N., Posthumus, H., Hubacek, K., Morris, J., Prell, C., Quinn, C. H., and Stringer, L. C. (2009). Who's in and why? a typology of stakeholder analysis methods for natural resource management. *Journal of Environmental Management*, 90:1933–1949.

Reis, J., Santo, P. E., and Melão, N. (2019). Artificial intelligence in government services: A systematic literature review. pages 241–252.

Rijksoverheid (2022a). Algoritmeregister voor de overheid. *Digitale Overheid*.

Rijksoverheid (2022b). Wet open overheid.

Rijksoverheid (2023a). Parkeervergunning aanvragen.

Rijksoverheid (2023b). Wanneer heb ik recht op bijzondere bijstand?

Rijksoverheid (2023c). Wetvoorstel: Wet open overheid.

Rodrigues, M. and Franco, M. (2021). Digital entrepreneurship in local government: Case study in municipality of fundão, portugal. *Sustainable Cities and Society*, 73.

Roehl, U. B. U. (2022). Understanding automated decision-making in the public sector: A classification of automated, administrative decision-making. *Service Automation in the Public Sector: Concepts, Empirical Examples and Challenges*, pages 35–64.

Rombach, D. and Steffens, P. (2009). Automation: What it means to us around the world. *Springer Handbook of Automation*, pages 1629–1641.

Roose, K. (2023). Don't ban chatgpt in schools. teach with it. *The New York Times*.

Rovers, E. (2022). Nu is het aan ons - oproep tot echte democratie.

Sambasivan, N. and Veeraraghavan, R. (2022). The deskilling of domain expertise in ai development. *Conference on Human Factors in Computing Systems - Proceedings*.

Sandeep Reddy, Sonia Allan, S. C. P. C. (2019). *Journal of the American Medical Informatics Association*, 27:491–497.

Satispi, E., Rajiani, I., Murod, M., and Andriansyah, A. (2023). Human resources information system (hris) to enhance civil servants' innovation outcomes: Compulsory or complimentary? *Administrative Sciences*, 13:32.

Schaefer, C., Lemmer, K., Kret, K. S., Ylinen, M., Mikalef, P., and Niehaves, B. (2021). Truth or dare? – how can we influence the adoption of artificial intelligence in municipalities? *53rd Hawaii International Conference on System Sciences (HICCS)*.

Schemmer, M., Kühl, N., and Satzger, G. (2021). Intelligent decision assistance versus automated decision-making: Enhancing knowledge work through explainable artificial intelligence.

Selten, F., Robeer, M., and Grimmelikhuijsen, S. (2023). 'just like i thought': Street-level bureaucrats trust ai recommendations if they confirm their professional judgment. *Public Administration Review*.

Shaw, J., Rudzicz, F., Jamieson, T., and Goldfarb, A. (2019). Artificial intelligence and the implementation challenge. *Journal of Medical Internet Research*, 21:e13659.

Sheridan, T. and Verplank, W. (1978). Human and computer control of undersea teleoperators.

Sheringham, J., Kuhn, I., and Burt, J. (2021). The use of experimental vignette studies to identify drivers of variations in the delivery of health care: a scoping review. *BMC Medical Research Methodology*, 21.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy.

Siau, K. and Kam, H. J. (2006). e-healthcare in abc county health department (abc-chd): Trade-offs analysis and evaluation. *Journal of Information Technology*, 21:66–71.

Siebert, L. C., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., Jonker, C. M., van den Hoven, J., Forster, D., and Lagendijk, R. L. (2022). Meaningful human control: actionable properties for ai system development. *AI and Ethics*.

Silva, A. S., Campos-Silva, W. L., Gouvea, M. A., and Farina, M. C. (2019). Vignettes: A data collection technique to handle the differential operation of items in surveys. *Brazilian Business Review*, 16:16–31.

Silveira, G. J. D. (2005). Improving trade-offs in manufacturing: Method and illustration. *International Journal of Production Economics*, 95:27–38.

Simonofski, A., Tombal, T., Terwangne, C. D., Willem, P., Frenay, B., and Janssen, M. (2022). Balancing fraud analytics with legal requirements: Governance practices and trade-offs in public administrations. *Data & Policy*, 4.

Sundar, S. (2023). If you still aren't sure what chatgpt is, this is your guide to the viral chatbot that everyone is talking about. *Insider*.

Tangi, L., van Noordt, C., Combetto, M., Gattwinkel, D., and Pignatelli, F. (2022). Ai watch. european landscape on the use of artificial intelligence by the public sector.

Terra, A., Riaz, H., Raizer, K., Hata, A., and Inam, R. (2020). Safety vs. efficiency: Ai-based risk mitigation in collaborative robotics. *2020 6th International Conference on Control, Automation and Robotics, ICCAR 2020*, pages 151–160.

Turing, A. (1950). Computing machinery and intelligence. *ind 49*, page 433 – 460.

Umbrello, S. (2020). Meaningful human control over smart home systems: A value sensitive design approach emerging technologies and value-based design methodologies view project posthumanism and ecophilosophy view project meaningful human control over smart home systems: A value sensitive design approach. *HUMANA.MENTE Journal of Philosophical Studies*, 13:40–65.

Umbrello, S. and van de Poel, I. (2021). Mapping value sensitive design onto ai for social good principles. *AI and Ethics*, 1:283–296.

UN (2020). Covid-19 pushing more government activities online despite persistent digital divide, annual e-government survey finds. *Technical Report*.

UN (2022). The sustainable development goals report.

van de Poel, I. (2021). Values and design. *The Routledge Handbook of the Philosophy of Engineering*, pages 300–314.

van de Poel, I. (2022). Understanding value change. *Prometheus*, 38.

van Engers, T. and de Vries, D. (2019). Governmental transparency in the era of artificial intelligence. *Legal knowledge and Information Systems*, pages 33–42.

Vinadio, T. B. D., van Noordt, C., del Castillo, C. V. A., and Avila, R. (2022). Artificial intelligence and digital transformation competencies for civil servants working group report on ai capacity building.

VNG (2020a). Artificiële intelligentie. *Trends Informatiesamenleving 2020*.

VNG (2020b). Vng vraagt meer aandacht voor uitvoerbaarheid van de wet op de artificiële intelligentie.

Wang, C., Zhu, X., Hong, J. C., and Zheng, D. (2019). Artificial intelligence in radiotherapy treatment planning: Present and future. *Technology in Cancer Research & Treatment*, 18:153303381987392.

Wickens, C. D. and Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science*, 8:201–212.

Wiesmüller, S., Fischer, N., Mehnert, W., and Ammon, S. (2023). Responsible ai adoption through private-sector governance. pages 111–132.

Wilson, C. (2022). Public engagement and ai: A values analysis of national strategies. *Government Information Quarterly*, 39:101652.

Wirtz, B. W., Langer, P. F., and Fenner, C. (2021). Artificial intelligence in the public sector - a research agenda. *International Journal of Public Administration*, 44:1103–1128.

Wirtz, B. W., Weyerer, J. C., and Geyer, C. (2019). Artificial intelligence and the public sector—applications and challenges. *International Journal of Public Administration*, 42:596–615.

Wirtz, B. W., Weyerer, J. C., and Sturm, B. J. (2020). The dark sides of artificial intelligence: An integrated ai governance framework for public administration. *International Journal of Public Administration*, 43:818–829.

Yakimova, Y. and Ojamo, J. (2023). Ai act: a step closer to the first rules on artificial intelligence.

Yigitcanlar, T., Agdas, D., and Degirmenci, K. (2022). Artificial intelligence in local governments: perceptions of city managers on prospects, constraints and choices. *AI and Society*.

Yin, R. K. (2014). *A (VERY) BRIEF REFRESHER ON THE CASE STUDY METHOD*.

Yoshioka, N., Husen, J. H., Tun, H. T., Chen, Z., Washizaki, H., and Fukazawa, Y. (2021). Landscape of requirements engineering for machine learning-based ai systems. *2021 28th Asia-Pacific Software Engineering Conference Workshops (APSEC Workshops)*, pages 5–8.

A APPENDIX D: VALUE SENSITIVE DESIGN

Hier VSD delen in plakken, kort stukje in hoofdstuk 3 noemen en dan verder hier ook stuk met de grandchallenges, dit ook kort noemen in hoofdstuk 3.

A.1 VALUE IDENTIFICATION THROUGH SUSTAINABLE DEVELOPMENT GOALS

When looking at the paper from Umbrello and van de Poel [2021] another framework, based on the VSD theory, has been suggested focusing on 4 phases where there is constant iteration between the phases and that overlap the other three overarching concepts from Umbrello [2020], as can be seen in figure A.1. This research will focus mainly on the first three phases where the results aim to provide guidance for the fourth phase where piloting, prototyping and experimenting can occur.

As Umbrello and van de Poel [2021] did, a start for value identification can be made by looking at the sustainable development goals (SDGs) by the United Nations and at the values that have been addressed in previous literature specifically on the use of AI [UN, 2022]. From these SDGs, one value arises in the implementation of AI in the public domain where equality and fairness would preferably be the result while respecting the level of human control and autonomy (SDG 10, Reduced Inequalities). But also the value of safety in innovation which could be considered as part of SDG 9 and 16, and responsible innovation which could be considered as part of SDG 9 and 12. These values are in line with some of the criteria from the previous chapter where human control and quality of service in terms of equality and fairness are considered the main criteria in determining a level of automation. The trade-off between these two is therefore extremely relevant to discuss.

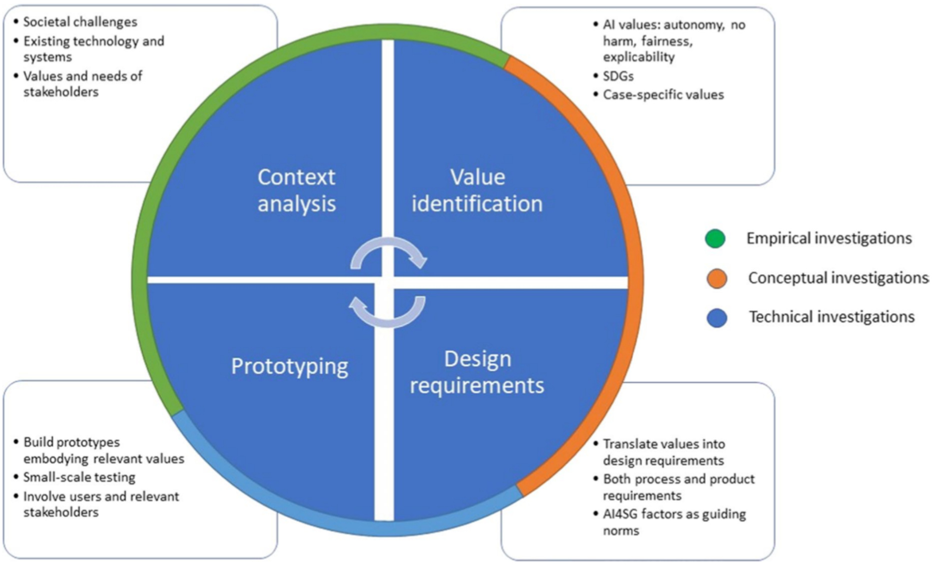

• Societal challenges
• Existing technology and systems
• Values and needs of stakeholders
• AI values: autonomy, no harm, fairness, explicability
• SDGs
• Case-specific values
Context analysis
Value identification
Prototyping
Design requirements
● Empirical investigations
● Conceptual investigations
● Technical investigations
• Build prototypes embodying relevant values
• Small-scale testing
• Involve users and relevant stakeholders
• Translate values into design requirements
• Both process and product requirements
• AI4SG factors as guiding norms

Figure A.1: VSD design process for AI technologies, adapted from Umbrello and van de Poel [2021]

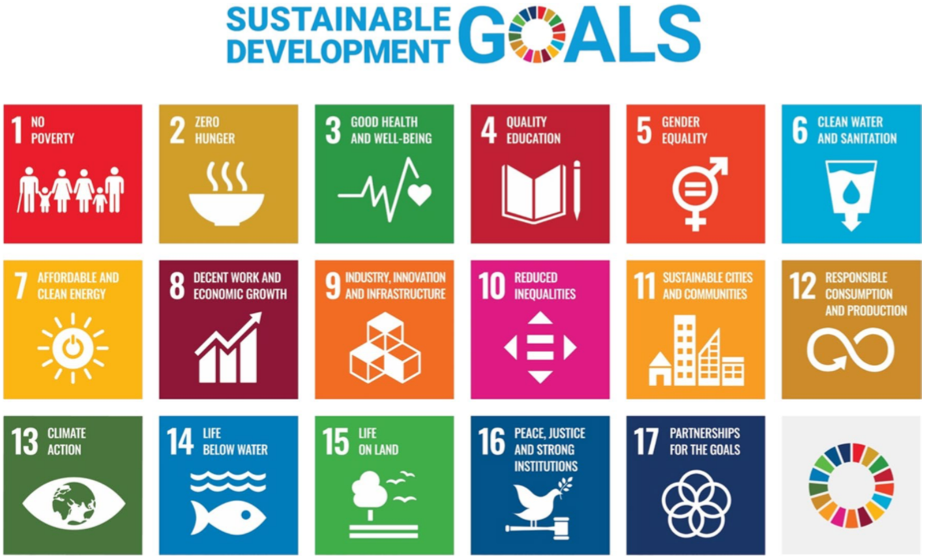

SUSTAINABLE DEVELOPMENT GOALS
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS

Figure A.2: United Nations Sustainable Development Goals, adapted from UN [2022]

B APPENDIX B: INTERVIEWS

This appendix includes the summarised transcripts with quotes from the expert interviews that have been held with 9 experts in total. Three interviews have been conducted with three different domain experts on the implementation of AI or algorithms in general in the public domain. These research domains are the public domain, the private domain and the scientific domain. The public domain eventually has to implement the level of automation and is also running pilots or has already implemented some forms of algorithmic decision-making and is therefore key to take into account. For this research, we consider public-private cooperation where the private domain provides an automation tool that can be implemented in the public domain. These private organisations (start-ups scale-ups) have been interviewed on their expertise and experience in implementing their tool in the public domain or how they would foresee this and what challenges and opportunities would arise. The scientific domain was included in this research to adhere to the three domains as also mentioned in the VSD model by Umbrello [2020] but also to provide more insights into the existing frameworks and theories that have been discussed in recent literature. The combination of these three domains will provide an insightful and multidisciplinary overview of the challenges and the risks of automation in the public domain but mainly on the criteria, context factors and role of the civil servant when deciding to automate decision-making of civil servants, where civil servant and an algorithm such as AI will together take a decision.

Table B.1: Interviewees Overview Appendix A

Interviewee	Domain	Role/Function
Interviewee 1	Science	PhD Researcher TU Delft
Interviewee 2	Science	Senior Researcher TU Delft
Interviewee 3	Public	Coordinating Policy Officer in the domain of AI
Interviewee 4	Private	Software Develop & Data Analyst
Interviewee 5	Private	Deep-Fake Expert
Interviewee 6	Private	Project-manager
Interviewee 7	Science	Researcher in Informatics and Responsible Implementation of Technologies
Interviewee 8	Public	Policy Advisor on Ethics
Interviewee 9	Public	Project-manager

In advance of every interview, the participants were asked for their function or role, next which could be used in this report to show what experts have been interviewed. They were also asked to indicate their prior knowledge of implementing AI in the public domain. Next to this, a brief explanation was provided where the concept of ”level of automation” was elaborated on as this could be confusing. Some participants only saw the use of algorithms as full automation and thus the replacement of civil servants. The level of automation was elaborated on when this was needed and the focus on the hybrid human-AI cooperation in these systems was made clear. Below all the interviews will follow, some of them in Dutch and some of them in English. Before these interviews are each shown, the main questions will be mentioned.

B.1 INTERVIEW PROTOCOL

The interview exists of three parts after the introduction questions: 1) Algorithmic decision-making, 2) Criteria in choosing a level of automation, and 3) the contextual factors that need to be addressed when looking at different levels of automation and having to determine a responsible level. The first section focuses on the following questions:

- *What would be the main advantage of using algorithms in the decision-making process by civil servants where hybrid cooperation between humans and algorithms (AI) is present? And elaborate on why this would be the main driving force.*
  - This question aims to understand the reasons why experts think automation would be beneficial in the public domain but most of them already mention the challenges and risks that also come with this
- *Should there always be a certain level of human control in decision-making in the public domain that has a direct impact on individual citizens?*
  - This question focuses already on the criteria of human control as it was found in the literature that this criterion is the hardest aspect to define and analyse. The role of the civil servant and the level of human control that is present should be very clearly stated before implementing any form of automation. This question already aims to get some more insights on this topic.

The second section focuses on the other criteria that are present and how they already view some trade-offs for different kinds of decision-making processes.

- *What criteria would or should play a role in determining a responsible level of automation for the decision-making by civil servants?*
  - This question aims to test whether the criteria found in the literature are indeed the same criteria that the experts have in mind. It is also used to see whether some criteria or more or less important in the public domain but also more specific in different decision-making processes.
- *What does meaningful human control entail? Or what should it entail?*
  - This question comes back to the level of human control but more specifically on the notion of meaningful human control as is often referred to in literature and where lots of debates exist. The respondents are asked to give their perception of what meaningful human control should entail

The final and third section focuses on the contextual factors that differ per decision-making process by civil servants that have a direct impact on individual citizens. This could for example be the sensitivity of a decision to be made, or the workload it gives to civil servants or maybe the impact that these decisions have on individual citizens.

- *What context factors play a role in automating the decision-making of civil servants? And how do these context factors impact the automation of decision-making?*
  - This question aims to find how different context factors have an impact on the responsible level of automation and could thus have a trade-off with the criteria as discussed in the previous section of the interview.
- *What extra knowledge do policymakers in the public domain need in order to determine a responsible level of automation for this kind of decision-making process by civil servants? Looking at good cooperation between humans and AI.*

  - This is a question to see how the role of the policymakers differs in the eyes of the experts from the role of for example the AI or the civil servant. The latter is the focus of the second question in this section.

- *What should be the role of the civil servant (the user of the algorithm) in using the algorithm in their decision-making process?*
  - This question adds to the previous questions of human control but focuses more on the role the civil servant should play, focusing on responsibility, autonomy, supervision, etc.

The interviews have been summarised and some quotes by the interviewees have been noted as these were found interesting for this research. Their answers are categorised into the three sections that the interview exists of as mentioned above.

B.2 INTERVIEW 1, PHD RESEARCHER TU DELFT

Algorithmic decision-making

1) De drijvende kracht is dat bij veel uitvoeringsorganisaties er gewerkt wordt met gegevens van miljoenen mensen. Dit is niet te doen zonder automatisering dan wel deels geautomatiseerd. Voorbeeld over Belastingdienst
2) Dan heb je algoritmes nodig waar ze verbetering kunnen brengen, hierbij moet het algoritme ervoor zorgen dat de mens zich kan richten op de taken waar echt aandacht voor nodig is. Het gestandaardiseerde werk haal je eruit en wordt gedaan door een algoritme.
3) Vaak zijn er best veel gegevens beschikbaar voor beslissers alleen is het lastig hier en volledig overzicht van te krijgen. Met behulp van een algoritme kun je verschillende variabelen wegen tot een soort eind oordeel en dat is in die zin heel objectief, omdat al die variabelen de hele tijd worden meegewogen. De data is niet objectief maar wel omdat alles mee kan worden genomen

"Binnen de publieke sector heb je altijd te weinig mensen om je werk goed te doen. En dat is heel logisch want je wordt betaald van belastinggeld en moet het constant met net iets te weinig doen". Daarbij moet je dus ook wel als publieke sector zeggen dat je de innovatie aangaat om taken op een betere en efficiëntere manier te doen. Waardoor het minder geld kost.

"menselijke controle is een van de moeilijkste dingen van het gebruik van een algoritme, dat is niet de technische kant, dat is allemaal wel te fixen. Menselijke tussenkomst wordt aan de ene kant vaak genoemd als iets wat echt nodig is, en aan de andere kant zie je dat het soms een tussenkomst wordt als een soort van vinkje, waardoor men het idee heeft dat er veel controle nog op is. Maar als iemand altijd het advies van een algoritme overneemt, dan is die controle er in praktijk eigenlijk helemaal niet."

Je raakt ook een deel van je kennis kwijt als je als ambtenaar telkens de beslissing of een deel van de beslissing in je schoot geworpen krijgt. Er is nog geen handige manier om met menselijke tussenkomst te werken. Hebben we nog wel invloed op algoritme? Dit zou constant ter discussie moeten staan.

Bepaalde taken is geen menselijke tussenkomst nodig. Bijvoorbeeld aangifte belasting. Maar het is wel echt nodig als jij als burger denkt dat er iets niet klopt dat er dan iemand is die menselijke tussenkomst biedt, waar jij vragen aan kan stellen en die uitleg kan geven. Dat is veel relevanter. "Contact met de mensen die tegen een probleem aanlopen met zo'n geautomatiseerd besluit is belangrijker." Vooral belangrijk bij massa beslissingen. Als het toch al over maatwerk gaat, v.b. reclassering, daarbij moet de reklasseringsmedewerker verschillende systemen naast elkaar houden en zelf de beslissing maken. Dat is veel meer een inschattingsvraag waarbij gegevens ondersteunen die weer gestructureerd kunnen worden door algoritmes. Hierbij is het kennisniveau van de ambtenaren dus heel belangrijk. Zijn ze instaat de uitkomsten uit die algoritmes te beoordelen en te wegen? Ander soort menselijke controle.

Criteria

Vanuit het algoritme bekijken. Als er beslisregels zijn die heel duidelijk de wet volgen, waar goeie registraties van zijn (Niveau van interpretatie). Dan stelt dit jou in staat om best goed geautomatiseerde besluiten te nemen. Sterker nog. Veel beter dan mensen. Dus dan hoge mate van automatisering logisch en mogelijk! Aan de andere kant, met wetten die veel maatwerk vergen, met veel interpretatie, en waarbij de data ook niet zoveel zegt over wat je nodig hebt bij de beslissing. En alles daartussen.

Als je hem vanaf de andere kant aanvliegt dan kun je zeggen dat bij de laatste categorie (met veel maatwerk) dat er bijvoorbeeld veel kennis nodig is en mag er weinig deskilling zijn. Is het meer de wet volgen, dan is deskilling niet zo erg omdat dit minder een expertise is en anderen dit ook zouden kunnen. "Wettelijke kennis kun je heel makkelijk programmeren in een computersysteem, terwijl bij andere zaken zoals bepalen van uithuisplaatsing van een kind, daar is echt een mens nodig om de context en de problematiek te snappen. Dan moet je ook veel meer weten van de uitwerking van een jeugdmaatregel." "Maatwerk vergt gewoon veel meer kennis". "Als je met een casus komt denk ik dat je adhv deze criteria heel goed kan zeggen welk niveau van automatiseren van toepassing zou moeten of kunnen zijn" Ze zijn allemaal erg belangrijk.

Voorbeeld belastingdienst (massa) efficiëntie en kwaliteit gaat hier echt omhoog bij automatiseren. Deskilling vindt natuurlijk ook wel plaats, maar dat is niet zo'n groot probleem. In deze casus weegt dit niet zo hard. In een veel gevoeligere casus, uithuisplaatsing, in principe kan dat veel efficiënter. En je zou ook nog kunnen beargumenteren dat een algoritme wellicht wel veel waarde-vrijere beslissingen maakt, met minder bias. Maar daar wil je toch niet dat de beslissing door een computer wordt genomen. Ik denk dat we vinden dat daar teveel bijzondere gevallen en uitzonderingen zijn, dat daar echt een mens naar moet kijken. Een mens kan hier beter mee omgaan en context in een scenario interpreteren. "Je kan een algoritme alleen maar meegeven wat je hebt aan data, en dat is vaak niet de context"

Acceptatie (van burger) is altijd heel belangrijk. Kwetsbaarheid is misschien minder belangrijk. Een makkelijker te automatiseren vraagstuk kan voor een kleine groep kwetsbare mensen heel pijnlijk zijn als dat verkeerd uitpakt. Maar er zitten zoveel kwetsbaarheden bij burgers waar de overheid rekening mee moet houden dat dit niet echt belangrijk is voor mate van automatiseren. Want dan is automatisering nergens nodig, "wat ook een heel slecht idee is". Bepaalde taken waar je voldoende kwalitatieve data voor hebt om besluiten mee te maken (vaak bij wet al geregeld) dan kan je hier veel efficiëntie winst boeken door te automatiseren maar als de context en de beslissing veel gevoeliger zijn (uithuisplaatsing), dan moet er iemand langs gaan om de sfeer te proeven en menselijke controle dus onmisbaar. Puur gericht op publieke sector. In andere sectoren (medisch) waar minder wetmatige beslissing zijn kun je wel ook veel halen uit beslisondersteuning Voor het automatiseren van beslistaken kun je naar verschillende onderdelen kijken, bijvoorbeeld het analyseren van data, het voorstellen van beslissingen en het nemen van een beslissing. Hoe ziet u automatisering in deze verschillende aspecten? Zelfs met advisering moet je heel erg oppassen! De laatste stap van het beslissen moet je aan de interpretatie van de beslismedewerker overlaten. Daar kun je veel meer op richten, hoe zo een persoon getraind is, etc.

Bekijken vanuit het algoritme en wat er al vastgelegd is in de wet. Vanuit de systeem veiligheidstheorie, elk systeem heeft een controller en een proces. Controller geeft input aan proces, en output uit proces kan de controller weer meten, zien en observeren. 4 condities voor veilige en betekenisvolle controle: 1) Controller moet een duidelijk doel hebben 2) Controller moet goed kunnen observeren wat de staat van het proces/systeem is. 3) Controller moet een model van het systeem hebben, weten hoe het werkt 4) Controller moet input kunnen geven (invloed hebben) Terug naar menselijke controle, moet kennis hebben, hoe output kunnen interpreteren, hoe invloed uit te oefenen en wat is het doel van de mens in dit systeem? "We leggen nu veel te veel nadruk als we het hebben over het werken met algoritmes, op het algoritme zelf. Waar we naartoe moeten is kijken naar de mens, en wat de mens moet kunnen om samen met het algoritme tot goeie beslissingen te komen. Veel mensen zijn totaal niet getraind in het omgaan met algoritmes. Hierin mis je dus punt 3 en 4, want je weet niet wat jouw input doet op het systeem en weet niet hoe het systeem tot een output komt.

Context

Gevoeligheid van de beslistaak en gegevens, impact op burgers, werklast van ambtenaren, wet- en regelgeving, enz. Belangrijk om context te definiëren. Mensen zitten in een bepaalde context welke vaak heel belangrijk is bij het automatiseren van beslissingen. In sommige gevallen is deze context niet belangrijk omdat alles wat je moet weten om een beslissing te maken al geregistreerd is. Totaal casus afhankelijk.
En je hebt de context van de ambtenarij zelf. Hoge werkdruk, kans op automatiseren zorgt voor ontlasten van druk dan is dit een reden om hiernaar te kijken. Maar als je ziet dat er een taak is waarbij veel kennis nodig is, waarbij ambtenaren veel beslissingen nemen welke niet te volgen is uit data. Dan is dat een hele relevante factor om het niet te doen. Je wilt toch dat de overheid goed met je geld om gaat dus een bepaalde efficiëntie verwacht je misschien wel. Dat is toch wel een waarde die men af en toe vergeet bij het automatiseren, de doeltreffendheid van de overheid. En basisbeginselen waarvan je mag verwachten dat de overheid deze altijd in het achterhoofd heeft: Gelijkheid, privacy (vooral in veiligheidsdomein, daarbuiten best wel "overhyped"). Belangrijk om te kijken naar "welke taken lenen zich nou echt goed voor automatiseren?". En dan kom je er vanzelf wel achter wat voor narigheden er dan op de loer liggen. En daar zijn de normale waardes van op toepassing. Dit mag je verwachten van een ambtenaar.

Beleidsmakers moeten echt goed snappen wat systeem doet en waar het systeem ook moeite mee heeft. Stukje vertrouwen zit hierbij. Voorbeeld van Chat GPT. De veiligheidstheorie die kun je hier heel goed op gebruiken.De ambtenaren vertellen het systeem wat het moet doen, die zijn ontwerper en hebben hierin verantwoordelijkheid in dat het systeem doet wat het moet doen. En bij de besluitvorming zelf moet de ambtenaar vooral in controle zijn van het systeem, dat het altijd weet wat het doet, en daar moet iemand op toezien. Dus hierin wordt de ambtenaar meer een toezichthouder dan een beslisser. En daar moet je voor getraind zijn.Mensen met veel kennis wil je in dienst houden: voor als het systeem faalt, of om burgers uitleg te geven over hoe zo'n systeem werkt en hoe je het kan gebruiken. Ook als doel op zich, het automatiseren van taken van ambtenaren moet voor meer contact tussen ambtenaar en burger zorgen. "meer tijd voor burgers zelf, voor de echte problematiek." Lastig als veel mensen ontslagen worden door automatisering, je hebt een basis nodig (politieke keuzes).

B.3 INTERVIEW 2, SENIOR RESEARCHER TU DELFT

Algorithmic decision-making

Imperial, efficiency, saving human work, productivity. Less discretion. "The driving force is efficiency, saving human work, productivity. As a tool of efficiency & productivity" Also potentially less arbitrariness in the situation, because of the standardization which is important in the public domain. Possibly corruption and arbitrariness on an individual level could be reduced.

Philosophically there always is human control. Whether someone has created the model, the data, or someone designing, data from human action that will affect the system. But the question is what kind of control and whether that control would be meaningful. "control is a very generic term, that covers various things. You could see control as influence, the less demanding sense of control. In that sense AI is always under human control because it is always influenced by human control, somehow. Whether by the designer, by the human data, by the way it is trained, by the interaction with people. At the other extreme there is control where one individual person, who has full knowledge of a specific system from a to z. From a technical component to the way it interacts, it is able to predict its behaviour. And that is control, that nobody ever has. And is not desirable. Because one of the goals of automating is not to burden humans with too much action. We want something in between. We want a form of control that is not just causal influence, but it doesn't require the human to do all the work. But we want a form of human control where there are relevant human beings, or groups of human beings that morally impact the outcome. To make for a responsible implementation and well implemented values, interest, procedures etc. Second condition is this group should be in such a relation with the system that they could be held responsible by others legitimately for the behaviour for the system." In terms of explainability and accountability. They could even be blamed to some extent if something goes wrong with harm to individuals. They should also be able to change the system, when the system is not good or not good enough anymore (update).

Criteria

"I like the idea of not looking at levels of automation per se, but defining levels of automation based on what you want to achieve with automation. Because you have the example of in cars, 5 levels of automation have often been defined based on the task of distribution. Level 0 is driver does it all, level 5 is car does it all. And the rest in between. But in doesn't tell you anything on the level of competence of the actors, the real level of meaningful human control of the actors, the distribution of morally legal responsibility. It just gives the idea of task distribution. But this is just really ideal. It doesn't really tell you whether and to what extent the differentiation will be actually able to do what they are supposed to do and to take responsibility for what they do." "So, in general my answer would be that, the level of automation should follow, and should be kind of defined, based on what level and meaningfulness of human control different agents are able to take. Rather than just, what would be the ideal distribution of tasks, based on some idealized agent. It should be more context dependent, it should be more directly and explicitly connected with the values we want to achieve with the responsibility people are able and willing to take. "Moral control moral responsibility is still central here"

"You should not automate more than people are willing to take, in terms of their capacity to maintain meaningful control and responsibility of the system." Criteria: efficiency and productivity is one, another one reduction of arbitrariness, bad discretionality, the kind discretionality that in the public domain is closer to corruption. Diversity and inclusion are super important. By definition systems tend to be optimized and standardized in one way that could fit well for certain users and not others. And that holds both on the side of the expert (public servant) but also on the citizens. And another hidden risk for automation is the implicit change of values (Importance of time) that you may have in a certain profession." There could be a big shift, because of other external people, ITS perks involved, from public values such as transparency, accountability, service to the citizens, towards the values of optimization, efficiency, you name it. This is a risk that you have everywhere. Because of the specific way the technology works there might be an implicit shift of values. This is connected with the shift of responsibility. One way this could happen, you have a shift in people who have a real agency in the system, in the case of ITS perks, as opposed to the public servant themselves. But also on the structure of the technology itself. The fact that the technology is learning, optimized because of its own nature. This is another part where the level of automation should be aligned with. A shift in internal goods and values of a certain interesting.

Yes human control is in philosophy an intrinsic value, so good in itself. But we should not overstate importance of meaningful human control. It may important to maintain in all of this applications. "It's a necessary but not so sufficient condition." It's a starting point. Context dependent how it is achieved, what do you want to achieve through MHC, is also dependent on different capabilities. "some say, once you reach human control, all ethical issues will be fixed. Check. But this is NOT the case." Issues of justice, fairness, inclusiveness. It is more likely to achieve all the other issues if you have meaningful human control, but it is not the holy grail. There are other dependent variables you have to consider.

Context

I would not frame it as levels of automation. But let's say error-sensitivity. Are we ready to accept a certain level of error-sensitivity. And a certain level of opacity are we willing to accept? These two things should be taken into account. By looking at for example how harmful can things be. One might say that those highly sensitive cases (matter of life death) need transparency, explainability, even financial matters could be very critical to people. So these also could be problems for low income person. And don't underplay social concepts such as trust in the system? If you start getting the feeling that the system is targeting you in an opaque way, even if it is a relative simple thing like a fine or parking tickets. This could really affect your wellbeing, trust in the system, etc. No simple answer. Roughly you may say that, depending on how impactful decisions can be on people's lives, then higher standards of meaningful human control and responsibility should be present. Which is prima feca is true, it is no caveat that this solves all the questions. Because sometimes even the more simpeler tasks where we would expect that less meaningful human control would be sufficient can still be very impactful on certain people's lifes. E.g. toeslagenaffaire.

How much error-sensitivity are we ready to accept? Opacity? How harmful can it be? Trust in the system by citizens. Should trust by civil servants also play a role? Those two perspectives are important. If civil servants don't trust the system that is a big issue. And this has to do a bit with taking responsibility of the system. So if the civil servants feel like they don't own the system, they will not feel responsible for it. They will probably disengage. There has been plenty of research on machine human interaction where even the simplest machines are not used or not accepted because there is a lack of trust. It might be controllable but if they perceive it not to be controllable, they will not trust it. Self fulfilling prophecy. Also from civil servant there should be trust. It should fit with the way civil servants work. Workload is important. There is a myth to bust here. People believe automation will reduce workload, but studies in psychology show that there are loads of overhead tasks that cover automation. This is true with every task. If you break down a task in different parts you have the illusion that this is more efficient. But there is always this hidden overhead in terms of coordination.

Context dependent. It is super important to have it clear, that all stages are potentially important. There is sometimes this simplified narrative that if you have one human somewhere in the chain that you would be fine in terms of meaningful human control. At one crucial moment this doesn't solve all the issues. If you have meaningful human control at the decision, e.g. driving a car, firing a rocket, taking a decision. But you were not there when they gathered the information on which you are taking the decision. Than still no meaningful human control, no understanding. But also when there is nobody to explain the decision, the chain of control is again messed up. This is exactly one way of defining human control. To look at the process from design of data acquisition, to the aftermath of the accountability process and to define who should be there doing what. To see that the whole system will deliver what we want it to deliver. Looking at the whole chain define who should be where doing what. Meaningful as in everywhere someone has to be present with certain value.

They need a level of technical knowledge obviously. In general in the public discourse, there is this idea of new skills that are needed. And usually what we have in mind is an update or a technological skill. And this is certainly part of the story BUT(!) social and soft skills and awareness of your new responsibilities are equally if not more important. This has to do with knowledge about your role, distribution of tasks, in this new system. Both from a technical but also a (moral) legal point of view. So knowledge on your moral and legal obligation. Knowledge about expectation of the other users, knowledge about the way you would be assessed by superiors. Knowledge is a lot bigger. It is a different organizational structure, you will need a lot of new knowledge about distribution of tasks, criteria of assessments of work, your relations with other (non-)human actors. A lot of this knowledge is not concrete knowledge on skills but more on know-how and motivational knowledge. You might know your role based on description, but you need to grow into your role. You need to trust it, believe it, and accept it.

It is the ultimate responsible actor in the system. Should be the responsible decision-maker. The person may not even be in the design or at the end pressing the button but the system is the designed that a human operator or group whose principles are reflected in the actions of the system and those are the ones who should take responsibility for the functioning of the system. As human actors in the system, civil servants should not be scapegoated, they should be made able to take responsibility in a meaningful way, otherwise it will not be fair to blame them if something goes wrong. Should be responsible for the output but also the procedures and transparency, explainability of these procedures. "it is a no brainer that in the public domain transparency is super important." In other domains it doesn't matter how something is achieved as soon as it is achieved (example of automated car that brings you somewhere

B.4 INTERVIEW 3, COORDINATING POLICY OFFICER IN THE DOMAIN OF AI

Algorithmic decision-making

Vaak waar veel data beschikbaar is, waar de financiële belangen groot zijn en ook waar het publieke domein een belangrijke functie heeft. Het wordt relatief veel in justitie en politie domein gebruikt, op het gebied van cold cases, handhaving, en veel in het toezicht domein, niet persé om fraude op te sporen. Ter besluitvorming, waar moet er onderhoudt gepleegd worden of opdroging van de bodem. Veel werken met ai in vorm van sensoren. Voorzichtige ontwikkeling als het om 1 op 1 besluitvorming gaat. Terwijl als je dan kijkt naar algoritmes wordt dat wel al heel lang toegepast, van belastingdienst tot svb. Zijn vaak wel simpele algoritmes, maar als het over AI gaat dan zijn mensen wel heel voorzichtig. Waar ai hier het meest gebruikt wordt nu is in de vorm van chatbots. Als het gaat om echte besluitvorming is ai veel vaker een hulpmiddel dan dat er directe besluitvorming uit voort vloeit. En als dat het geval is komt er vaak nog een check achteraf. 1 op1 besluitvorming: waar je het belang van een enkele burger raakt. Daar moet je wel heel transparant in zijn en goed zijn in het gebruiken van AI. Die bewustwording is veel groter geworden.

Je moet een hele grote afweging ook maken naar de consequenties! Sommige besluiten hebben minder grote consequenties en vereisen daarom minder menselijke controle. En ten tweede moet de feedbackloop hier wel heel erg goed zijn. Van burgers terug naar publieke instanties, om bijvoorbeeld in beroep te gaan of om te vragen om een uitleg (explainability). Goed ingebed zijn in de organisatie. Transparantie is hier superbelangrijk. Check ups and feedback loops moeten goed op orde zijn. "ik denk dat het wel onderschat wordt, dat AI heel vaak een hulpmiddel zal blijven. Dat blijft een hele belangrijke functie." Dan zul je zien dat nog steeds de persoon een laatste woord heeft, alleen wordt deze veel meer gesteund in zijn besluitvorming. Die wordt veel meer voorbereid op en veel meer geselecteerd. Net als bij de automatische piloot er moet een mens bij zitten voor check up voor onverwachte omstandigheden. En in een aantal gevallen zal dit een hulpmiddel blijven omdat deze meer patronen aangeeft ipv dat deze er resultaten aan geeft waar die wel. Waar die wel besluiten geeft kun je het doen, en kijk je vooral hoeveel besluiten neem je per jaar. Hoe groot neem je de foutmarge (risico impact).

Criteria

Simpelere en minder complexe taken kunnen veel makkelijker geautomatiseerd worden waarbij efficientie winst het meest belangrijke is. Ook sneller en acceptatie is hierin ook ontzettend belangrijk aangezien men simpele beslissingen graag snel genomen ziet worden. "Je gaat geen 100 jaar doen over een parkeervergunning" Kwetsbaar en deskilling zijn hier niet belangrijk in want men is liever bezig met het probleem van betere parkeerplaatsen regelen dan daadwerkelijk het behandelen van de vergunningaanvragen. Dus automatiseren van simpelere taken zorgt ervoor dat ambtenaren meer tijd hebben voor taken die belangrijker zijn, menselijker zijn, meer toevoegen en voor hen zelf ook als belangrijker voelen. Menselijke controle gaat er vooral om dat de burger in beroep moet kunnen gaan en aan moet kunnen geven als er iets mis is gegaan.

Complexere en gevoeligere beslissing staat kwaliteit echt voorop want juist in gevoeligere zaken kan het automatiseren de kwaliteit verhogen en het ook harmonischer kan maken. Ambtenaren nemen bijvoorbeeld verschillende beslissingen voor dezelfde scenarios maar afhankelijk van de situatie en het moment waar zij zelf in zitten ook. Dus hierin kan automatiseren de kwaliteit en consistentie verbeteren. Kwetsbaarheid is dan ook belangrijk. Goed uitleggen, transparant maken van beslissing, feedback loop moet zo snel mogelijk zijn, waarmee dus de menselijke controle ook ontzettend belangrijk is (ook vanaf de burger de menselijke controle. Meaningful human control ligt ook een groot aspect bij de burger. Ben je het er niet mee eens? Ga in beroep, dit moet makkelijk kunnen. De overheid moet makkelijk kunnen faciliteren dat de burger in beroep kan gaan en snel in contact komt met een ambtenaar. Efficientie en deskilling zijn dan minder belangrijk. Maar efficientie wel iets belangrijker dan deskilling. "Deskilling is juist zo, als je op dit soort moeilijke (complexe en gevoelige) dossiers geholpen wordt door AI, dan krijg je juist meer ruimte voor die menselijke maat. En dus om je echte skills in te zetten in plaats van dat je helemaal verdwaald in heb ik het wel goed gedaan. Lees uithuisplaatsing bijvoorbeeld." Ik kan met niemand vergelijken, ik kan niet inzien wat collega's gedaan hebben. En dat men zich hier ongelukkig onder voelt.

Stel een beslissing op niet een hele grote schaal: Het moet de menselijke maat omvatten. Bijvoorbeeld in het kader van gemeentelijke besluitvorming, soms kan een besluit rechttoe rechtaan direct kloppen, maar vanuit het menselijke oogpunt niet. Dan zijn de omstandigheden dusdanig dat je dit niet hebt kunnen zien of meten met het algoritme (context). Dat soort controle mechanismes moeten er wel in zitten op de 1 op 1 besluitvroming. Dat je de logica ook behandeld. Doe je dit dan altijd? Of steekproef gewijs? Op basis van signalen? Gebruikt het ai maar wil je wel het complete beeld aanspreken.

Context

1 op 1 besluitvorming: Wat zijn de consequenties voor de burger zelf, kun je het makkelijk corrigeren? Is het groepsgewijs? Toetsing op causale verbanden is belangrijk. Dat je bijvoorbeeld ai preventief wil toepassen (is de kans in Feyenoord groter op fraude of criminaliteit.) Dan is een toets heel erg belangrijk in de zin, welk label plak ik hier op, wat doe ik hiermee. Wat veroorzaak ik hiermee, hoe absoluut gebruik ik dat. Gaat er meer over wat betekend zo'n label op termijn voor mensen. Menselijke maat komt hier op een andere manier naar voren dan bij 1 op 1.

Bij 1 op 1 is het nut duidelijk maar bij groepen minder. Zijn de acties die ik neem echt verantwoord en nodig? Als je zegt dat het ondersteund voor het maken van beslissingen, dan ga je heel sterk kijken naar hoe waardevol een hulpmiddel is. En wat dan nog wel eens vergeten wordt, is dat ai het niet over gaat nemen maar dat het juist de kwaliteit verbeterd. Want deze persoon houdt meer tijd over voor meer op maat zorg en betere kwaliteit van dienstverlening. Dan komt het veel meer aan de orde of iemand goed genoeg geschoold is, Wordt het op de goeie manier gebruikt, gaat hij/zij het gebruiken, is het goed in processen ingebed, vind de organisatie het normaal dat je dit gebruikt? Dan komt het organisatorische aspect naar boven.

Heel goed weten wat de consequenties zijn en waar is het systeem op gebaseerd. Zelfs als iets gefundeerd kan zijn moet hier controle op zitten in gedachten houdend de context en de situatie. Dus belangrijkste is de kennis over consequenties en ten tweede waar is het op gebaseerd? Hoe verfijnd is het model? Toeslagenaffaire waren niet verfijnde schema's die gebruikt werden. Transparantie is hier ontzettend belangrijk. Misschien minder over het echte algoritme, maar wel over welke algoritmes zijn gebruikt, welke kennis en hoe men tot een besluit is gekomen. Maatwerk moet wel altijd toegevoegd kunnen worden. Dit gaat op basis van bijvoorbeeld 100 gevallen 99 gaan goed, 1 niet, dat je als ambtenaar je dan vrij voelt om te zeggen hey deze klopt niet! Tijd is enorm belangrijk om mee te nemen, dit verandert heel snel dus updaten moet ook beter en sneller kunnen. We denken vaak nog oud en traag terwijl systemen sneller kunnen veranderen. Dit vraagt om controles vanuit de mens. En in systemen inbouwen! Met verstand van zaken hiernaar blijven kijken. Onze tijd vraagt snelheid van handelen. En niet gaan wachten tot mensen gaan protesteren op het Binnenhof.

Blijven communiceren met de techniek. Werkt het wel of niet. Goed inbedden in proces. Deelverantwoordelijkheid hier ook voor ambtenaar. Helpen om collega's in te werken en het te begrijpen. Kennis doorgeven over ai gebruik, berijpen waarom en wat je doet.

B.5 INTERVIEW 4, SOFTWARE DEVELOP DATA ANALYST

Algorithmic decision-making

It makes subjectivity explicit, especially in the public domain where you are dealing with cases with sensitivity and where you act with your gut feeling. It is important that this gut feeling is made explicit, such that it can also be transparent to the clients and maybe even to the higher government when they inspect why they make certain decisions. Making explicit that it is pretty dependent on the person just as one decision maker but as an organisation that works in the public domain they might trade-off certain criteria with some subtlety to it. There is no hard rule that you should consider for instance the financial situation compared to the family situation twice as hard as the first one. There's no hard rule, but you need to address them both. So how do you as an organization weigh of those criteria, it is not yet clear anywhere. But you do have a certain approach, making that explicit and transparent is beneficial for them.

I think there should always be human control, because I think a model is just a representation of the data that you have. E.g. garbage in garbage out. But sometimes you don't even know if it is garbage. So It is very important that what is outputted is always monitored such that you see whether it is garbage or not! This is the task of the human. And it's just a tool. It's not autonomous in itself. It's just something you use.

Criteria

The impact of the decision, it is a general thing and can be cost or lives. That is very contextual. Impact first, then quality because quality can be deceptive as well. Circumstances can change so the context in which you make the decisions can change while the model stays the same. The model might be built good on the context it was built in. But if the context changes than you don't see it whether the model is still good. Also data selection error, you trust your model on data that is not represent on the real situation. May seem good but is not good. So it is very hard to find an objective measure of model so I wouldn't say that should be the first criteria to decide on.

Deskilling, especially with very simple decisions this can happen, but how bad is that?? Humans must still be a part of decision-making. If the humans were on the first place already hired to do these positions, of course it depends on how complex the decisions are. Doing some calculations for example is not really a big loss. For an AI to be an AI there should be some intelligence that you replicate that you create artificially. That is the remains of the civil servants themselves, They cannot be that easily replaced either. You could also see it, as instead of, removing or decreasing the numbers you reduce their workload on simpler tasks and maybe if you can train them to do more complex ones.

Humans should always be able to inspect the model. So how the models are making decisions. And the humans should always be able to give Feedback to change the model. That's basically what transparent means. That's how humans can control the model by knowing what the model is. Otherwise they don't know what they control.

Implementation effort, acceptability. They might be overlapping. The harder it is to accept the higher the effort for implementation (trade-off). If I would take acceptability out of the effort of implementation it would be the time to train civil servants. I don't think that is a very big criteria, if the efficiency adds up to the effort of implementation. If training civil servants is a huge tasks, this wouldn't be such a big deal, if the efficiency increases a lot as well.

Context

Workload, If they already have a very high workload it would be more interesting for them to get some help to automate things. But not fully automate things I would say. Not fully automate but would be more beneficial to automate to a certain extent. I also notice that staff or employees don't want to get rid of their complex tasks, and they still want to have the easy ones. It is not really about the workload but about the balanced workload Acceptability? Always working on the more complex tasks is not valuable, so creating a nice balance in workload of simpler and more complex tasks. Not to minimize workload.

Emotional burden is another aspect, so then effect of making wrong decisions on the civil servant. Feeling shame or blame. Responsibility. IF you are deciding on something impact full on an individual, this person may feel very guilty for having to decide this. It affects you as a decision maker. It has a big impact on you so also on the level of automation. You might want to get some help. If you completely automate the decision than no one will feel responsible for it. Therefore you shouldn't want full automation. In any level before that, each single decision can be partially automated but no single decision should be made fully automated. So for example 4 out of 10 cases is automated and 6 not to not lose the skill and to have efficiency gain as well is NOT a solution! Interesting idea nonetheless.

Automating very sensitive decisions could help in releasing the emotional burden because at least the civil servant will have some support in deciding this case. So it's like if you have to answer the question and look it up first. It also depends on the AI. But for example with expert input the AI is giving a result based on your fellow colleagues and experts. So a civil servant can feel supported in his or her decision. So yes, it can offer a help on the reduction of emotional burden, but I think it's also very personal. So it might but it might not.

Every actor should maintain the responsibility on the final decision that they make. Awareness on how the ai affects their decision. At some point they might not realise, that they have actually be fully dependent on it. So be aware of the affect of using AI even as a decision support tool. They may still feel like they are making the full decision while they do not realise that actually they have been following the AI. So a level of introspection would be wishful here.

B.6 INTERVIEW 5, DEEP-FAKE EXPERT

Algorithmic decision-making

AI is heel erg snel, dus daarmee draagt het bij aan efficiëntie. Dus ook zeker als je veel besluiten moeten maken. Ook is AI 24/7 bezig, en kan het beslissing nemen. Als je de AI goed implementeert dan kan het ook bias voorkomen. Soms heb je menselijke bias en als je de AI op de juiste manier implementeert dan kan je menselijke bias boven water krijgen of zelfs zien. Drijvende kracht voor het gebruik van AI is de hoeveelheid data. AI kan hier een oplossing in zijn. Kan hier makkelijk snel doorheen met een goed antwoord.

Zolang het mogelijk is, is het wel beter dat er een mens meekijkt met wat een AI doet. Met name wanneer de gevolgen hiervan groot zijn. Bijvoorbeeld het afkeuren van een persoon op iets. Dan is een menselijke controle erg handig. Waarom wordt deze persoon afgekeurd? Je kan hem ook breder trekken. Misschien moet wel duidelijk zijn waarom AI deze persoon afkeurt. AI kan nog best wel eens een blackbox zijn. Explainable AI bestaat wel maar is ingewikkeld en vereist ook weer een goede implementatie. Als je het goed, netjes, uitlegbaar doet dan kun je ook naar boven halen met welke features en welke biases de ai keuzes maakt. Dan kunnen mensen ook beslissen of ze het er mee eens zijn dat er hierop of daarop een keuze gebaseerd wordt.

Criteria

De gevolgen van de keuzes die gemaakt worden, bijvoorbeeld de bankrekening, als AI jou afkeurt voor een bank, dan is dit wellicht best wel een groot gevolg. Dus gevolg en impact is het grootst wat zou moeten meespelen bij de keuze om iets te automatiseren. Impact, gevolg. Hoe kritisch is het? Heeft het grote gevolgen op mensen laten we dan niet meteen beginnen met automatiseren, eerst kleine stapjes! Systeem ook testen. Doet de AI hetzelfde als de mens of wellicht zelfs beter? Dit moet je gaan testen. Als hij het beter doet kun je wellicht wel verder gaan automatiseren, doet hij het slechter, dan moet je het niet automatiseren. Met kanttekening van explainability. Zelfs als hij het goed doet moet het uitlegbaar zijn. Waarop baseert de AI keuzes? Dat is een ingewikkeld verhaal, maar dat kan je wel naar boven halen. De flaws en kanttekening van je AI moet je kennen en weten en begrijpen!

Acceptatie van burgers en ambtenaren kun je niet uitsluiten en mensen moeten er wellicht nog een beetje aan wennen dus daarom niet zo super belangrijk. Soms kan AI het gewoon beter. Maar als jij iedereen boos maakt door iets te automatiseren krijgen we dan een betere maatschappij?? Nee... Dient het AI de maatschappij?? Belangrijke vraag.

Met name dat er nog een mens bij staat om dingen te kunnen doen. Des te meer impact er is des te belangrijker dan er een mens bij de ai is. Daarbij geldt ook weer hoe goed ken je het systeem werkelijk. V.b. auto. Meeste mensen weten niet hoe een auto werkt, toch stappen we er ieder dag in en vertrouwen we dit apparaat met ons leven. Kunnen we dit ook hetzelfde doen met AI? Belangrijke verschil hier is dat desondanks dat jij de auto niet begrijpt zijn er genoeg mensen die dit wel snappen en die de auto veilig kunnen maken. Met Ai is dat net ietsjes ingewikkelder. Je kan naar een AI expert gaan maar omdat het toch wel als een blackbox is kan deze expert niet meteen zien of deze AI veilig, verantwoord en goed is. Als jij AI explainable hebt gemaakt, weet op welke data deze is getraind, weet of er biases zitten, en je bent een goed ingenieur, dan kun je langzaam zeggen. Oké jij begrijpt je ai zo goed, dat als jij zegt dat hij veilig is dan zouden we potentieel naar volledig automatiseren toe kunnen. Maar alleen dus als je dus daadwerkelijk dat niveau van begreep hebt gekregen. En wat gebeurd er als het mis gaat? Welke veiligheidsmeasures. Een auto zit er vol mee, stevig frame, airbag, riemen, etc.

Context

Als een AI een beslissing maakt die over het kind gaat dan wil je een mens erbij hebben ook voor de uitleg aan het kind. Het kan zo explainable mogelijk zijn maar je zult een mens nodig hebben om het emotioneel over te brengen aan het kind. Twijfel of een AI ooit dat laatste zou kunnen. Omdat empathie belangrijk is voor het maken van een beslissing. Heel belangrijk om de grenzen van je AI te kennen, Context kan meegenomen worden als de AI hier dusdanig in is getraind. Als er oorlog of covid is dan kan de ai getraind zijn om dit mee te nemen in de beslissing en dus eventueel milder te zijn. Maar kost meer moeite met training etc. en hoe gewenst is dit. KEN de grenzen van je ai. Voor nu lijkt het best wel zo dat AI's niet goed zijn in het begrijpen van context. Dus als context heel erg belangrijk is voor het maken van de keuze. Dan is een niveau van automatiseren minder mogelijk en of gewenst. Maar dit zal nog wel veranderen. Het gesprek is daarmee ook super speculatief. Wat kan er nu, waar gaan we heen? Wat kan er straks

Als een beslissing heel vaak langskomt, dan klinkt dit als een van de pijlpunten waar AI heel goed kan inspringen. Efficiëntie . Neemt de AI banen af, omdat er minder mensen nodig zijn. Dan is het antwoord waarschijnlijk ja. Aan de andere kant zie je ook dat er veel tekorten zijn in werkgebieden, heel veel ambtenaren zijn overwerkt dus hierin zou AI juist weer goed kunnen bijdragen. Door dingen te doen waar de mens vaak geen tijd voor heeft. Waar je misschien nog wel voor zou kunnen uitkijken is wanneer je ondersteunende AI hebt, Dus een AI die een ambtenaar ondersteund op bepaalde taken. Desondanks de controle die erbij is, dat er dan nog steeds een kans bestaat dat de AI de mens adviseert op dingen die hij/zij anders misschien niet zou doen. Voorbeeld militaire domein automatische wapens, advies is er al en insinueert al dat men moet schieten dus de mens zal sws sneller schieten. We moeten opletten op al deze best practices, en ook opletten dat desondanks dat de ambtenaar nog meekijkt of niet, zien wij eigenlijk een nieuwe bias ontstaan? Door te snel naar AI te luisteren. Dat zou best wel eens kunnen.

De eindverantwoordelijke is altijd de persoon die de AI heeft gekozen. Die gezegd heeft, ik gebruik voor dit proces, deze AI. De persoon die met de AI werkt, heeft ook een gedeeltelijke verantwoordelijkheid. Dus de verantwoordelijkheid splits zich op bij hybride vorm tussen mens en AI. De bouwer van de AI heeft ook verantwoordelijkheid, want deze adviseert waar de ai op gebruikt kan worden. Als dit misgaat heeft deze partij hier wellicht een aandeel in gehad. Klant moet goed ingelicht zijn over de mogelijkheden en de grenzen van de AI. De AI is nooit verantwoordelijk!!

B.7 INTERVIEW 6, PROJECT-MANAGER

Algorithmic decision-making

Met grote hoeveelheden beslissingen, zeker voor de overheid zelf, dus wat minder voor kleinere gemeenten kun je efficiëntie winst behalen. efficiëntie. De simpelere zaken waarbij er geen afweging nodig is van de medewerker maar welke echt op basis van beleid is. Als je deze kan automatiseren dan kan dat heel waardevol zijn. Zou ook kunnen bijdragen aan Consistentie van besluitvorming. Waarbij je lastige zaken kan uitlichten. Automatisch identificeren van lastige zaken om deze te overleggen. Stukje efficientie en kwaliteit winst is mogelijk mits toegepast op de juiste manier.

Als je een vergunningaanvraag hebt die echt op regels wordt afgedaan door een medewerker dan kun je deze ook automatiseren en dan veel sneller. Als je een zaak hebt waarbij de expertise van een expert (ambtenaar, medewerker) echt nodig is, dan vraag ik me af of dat te automatiseren valt en denk dat je dan altijd wel een behandelaar of medewerker nodig hebt die er dan nog naar kijkt. Bijvoorbeeld met ondersteuning van AI dan. Maar daar wordt automatiseren lastiger. Ook als de uitkomst direct invloed heeft op een burger, want dan maken burgers toch echt eerder fouten dan zo'n systeem. Moet echt getest worden voordat het geautomatiseerd wordt. En als je dan ziet dat burgers eigenlijk even vaak dezelfde fouten maken dan kun je het voor de maatschappij gewoon automatiseren, want efficiëntie winst. Voor iedereen beter denk ik.

Criteria

Efficiëntie, kwaliteit maar ook de technische capaciteit van die systemen. Hoe moeilijk is het om zo'n beslissing te vangen? Zijn het regelsystemen tot menselijke afwegingen (die wij dus kunnen automatiseren) maar daarin zit nog wel een verschil. De impact van een keuze speelt een rol, stel dat echt de zwakkere van de samenleving bij een foute beslissing, extremen gevolgen kennen dan is dit zeker niet wenselijk. Huiveriger om te automatiseren dan.

Acceptatie van ambtenaren is lastig, het introduceren van zo een systeem/model is verreweg makkelijker als je de ambtenaren meekrijgt die er mee moeten gaan werken. Bij veel toekomstige gebruikers is er echter de angst dat zij hun baan gaan verliezen. Dus daar hangt de acceptatie van af. En ook wel uit een stukje verantwoordelijkheid voor hun werk. De ambtenaren willen het vooral goed doen, en goed dun voor de burgers en zij vragen zich ook af of systemen dit even goed kunnen als dat zij dit kunnen. "Dus het is een stukje angst om hun baan te verliezen maar ook angst dat de kwaliteit van dienstverlening verlaagt." Zorgen over de kwetsbaarheid van burgers.

Dus misschien ligt de menselijke controle wel meer bij de beleidsmaker van bij de ambtenaar die met een algoritme gaat werken. Het management moet goed nadenken hoe ze zoiets willen introduceren. Bij acceptatie burger is het vooral angst, zeker met toeslagenaffaire. Zulke systemen worden gezien als eng, foutief of niet begrepen. Daarin is transparantie en uitlegbaarheid heel belangrijk want je moet een burger kunnen uitleggen, dit is de reden. Daar moet een mens tussen zitten. Dat is menselijker, zeker bij een overheidsdienst, dat wil je niet door een model of chat GPT laten doen. Iemand moet dit snappen en de burger te woord kunnen staan.

Ik hoop dat als je dingen efficiënter maakt, zoals simpele besluiten, die niet door ambtenaren genomen hoeven te worden waardoor zij meer tijd hebben om hun aandacht te richten op casussen waar hun expertise ligt. Dan kan ik me voorstellen dat er wellicht vervangen tot een bepaalde hoogte nodig is maar je hebt altijd ambtenaren nodig om de beslissingen toe te kunnen lichten en om feedback te kunnen geven aan het model. Service gerichter wellicht dan het zelf beslissen. Meer tijd en ruimte om met burger in gesprek te gaan waarbij transparantie en uitlegbaarheid heel belangrijk zijn.

Voor de komende jaren gaat er een deel zijn wat (nog) niet geautomatiseerd kan worden. Die ambtenaren wil je behouden! Je wilt voor die complexere taken dus ook personeel behouden! Neem ze mee in het hele traject over besluitvorming over algoritmes en stop er niet mee als want dan kom je nergens, maar leg ze wel uit, dit gaan we doen om dezen redenen, hier hebben wij jou nodig en je bent hier hartstikke waardevol. Doe het samen. Medewerkers waarderen het ontzettend dat ze meegenomen worden en dat er geluisterd wordt naar wat zij vinden en dat ze feedback geven op modellen.

Context

Impact speelt een rol. Een beslissing met minder impact is makkelijker te automatiseren, dan een beslissing die ontzettend veel impact zou kunnen hebben. Ook de 'ripple effects'. Dus niet alleen de beslissing zelf, maar ook de consequenties die vast zitten aan een beslissing die pas later boven water komen. Dat hele proces moet je in kaart brengen, het volledige gevolg van een besluit om een individu in kaart brengen. Technische mogelijkheden. Vaak willen we automatiseren maar zien we dat de input waarden nog helemaal niet te automatiseren zijn. Vaak worden de inputwaarden voor het automatiseren van een beslissing nog ingevuld door de medewerker, dus hierin ben je afhankelijk van de medewerker, (bias??). Als je alles wil automatiseren moet je dat deel ook automatiseren. Daar lopen ze vaak tegen 2 dingen aan: 1) het opstellen van die criteria is lastig, want die zijn niet digitaal beschikbaar, Je hebt een persoon nodig die er soep van maakt. Subjectieve afwegingen (urgentie bijvoorbeeld) Is dit technisch en conceptueel te automatiseren?

Beginnen met lager niveau van automatiseren, ook als de beslissing niet complex of lastig lijkt. Om te zorgen dat de kwaliteit van het model goed genoeg is en dat je dus niet regels introduceert waarbij je achteraf merkt dat het toch niet werkt in de praktijk (toeslagenaffaire). Altijd eerst testen, om te kijken of die regels echt de werkelijkheid vangen, of dat er toch dingen zijn die aangescherpt moeten worden.

Ook verschil tussen algoritme vanuit overheid zelf of externe partij. Binnen de overheid werken ze met data van burgers dus hier moeten ze heel erg voorzichtig in zijn. Voorbeelden gehoord over lange contracten met providers die vervolgens bepalen waar het wordt toegepast en waar niet. Super afhankelijk hier van een externe partij. Ondanks dat het intern ontwikkelen van software/applicatie langer duurt kan dit misschien wel veiliger zijn waarin het makkelijker toe te passen is. Samenwerking met externe partijen kan hier wel. Maar je moet de ownership van de modellen behouden, moet bij de overheid liggen. Vaak wurgcontracten met IT suppliers.

Als je niet voor volledig automatiseren kiest, dan zou ik zeggen ambtenaar, volg het systeem wees kritisch, wat zijn de criteria. Dus hier ook transparantie in het systeem. Ben je het er mee eens. Zo niet, beargumenteer waarom niet. Zorg ervoor dat ambtenaren meer de tijd hebben om echt bezig te zijn met de impact van de beslissing en de redenen waarom, dan het maken van de beslissing misschien wel. Geeft dus hier heel duidelijk winst in kwaliteit van de dienstverlening aan de burgers.

B.8 INTERVIEW 7, RESEARCHER IN INFORMATICS AND RESPONSIBLE IMPLEMENTATION OF TECHNOLOGIES

Algorithmic decision-making

Het aanbrengen van een zekere mate van consistentie. Die publiek goed overdraagbaar is. De recept metafoor. Je spelt uit wat je basisrecept is! Goed/beste/enige recept doet er niet toe. De ambiguïteit in recepten (koken) is ook goed. Iets opschrijven en daarmee dus accountability mechanism inbouwen. Je zegt wel dat je het zo doet, maar je doet het zo. Spiegel voorhouden! Interne spiegel. De algoritmische stap maakt dit concreter dan als de mens zoiets gaat doen. Een meer systematische verantwoording afleggen (mbv algoritmes).

Hangt af van je beleid. Maar consistentie waarschijnlijk omdat je een soort gelijkheidsprincipe wilt inbouwen. In heel veel praktische situaties moet je onder schaarste iets gaan kiezen. En dat betekent dat je bepaalde mensen wel een kans geeft en anderen niet. Als dat zo is, en iedereen is "gelijk" dan is de vraag: wat betekent dat gelijk zijn en kunnen we daar een gevoel bij en kunnen we daar iets op borgen of niet? En dat is tot nu toe een vaag concept geweest, ook op basis van de WOB-documenten die de respondent heeft mogen inzien. Vaak gut-feeling afgehandeld. Terwijl daarover kunnen spreken en legitieme afwegingen kunnen maken. Dan moet je even iets beter uitspelen wat bepaalde standpunt echt betekent. En of het contrasteert en of het woord wat je er aan verbindt, daadwerkelijk die uitwerking heeft. Qua bijvoorbeeld wat is eerlijk of rechtvaardig? Minstens 20 definities die elkaar wiskundig tegenspreken. Het helpt dan om te zeggen, ik ben eerlijk, dit is het recept waarop ik dacht dat ik eerlijk was. Van te voren dus vastgelegd. En dat iemand anders dan kan zeggen "dat is helemaal niet eerlijk, want xyz."

Criteria

In het ideaal beeld , hebben zowel ambtenaar als het algoritme, complementaire kwaliteiten en vaardigheden. Het algoritme kan super consistent iets gaan doen. De ambtenaar is een mens, dus dat is diegene die de menselijke maat er vervolgens tegen aan kan zetten. In de ideale wereld houden die 2 elkaar in balans en houden zij elkaar accountable. Dus zou zo'n algoritmische stap kunnen helpen om zo een persoon misschien op te schalen of breder te kijken , maar juist kan het complementeren in de dingen waar de persoon niet zo goed in is.

Low level taken zijn goed gekaderd, dingen die je kunt systematiseren of opschalen, daar kan je het (algoritme/AI) goed voor inzetten. Of als je dingen wel kan formaliseren, dus als het niet makkelijk te systematiseren is maar waarbij er wel heel duidelijk outliers zijn of patronen, als je dat wel in een systematisch proces kan kaderen dan kun je ook hier automatiseren. Zodra je een spiegel voor kan houden voor de ambtenaren die beslissingen nemen met soortgelijke beslissingen waarbij ze anders hebben gehandeld, dan is het gerechtvaardigd om hier een aantal stappen uit te automatiseren.

Kwaliteit is inderdaad het doel "maar dan wel goed gekaderd in de werkelijk impact die je wilt bereiken. Kwaliteit heeft natuurlijk heel veel verschillende interpretaties en in de AI wereld is dat meestal accuracy of loss. Dat is niet noodzakelijk, denk ik." De KPI's in publieke sector zijn toch vaak een gereduceerde versie van wat je werkelijk wil bereiken.

"Meer een standpunt in nemen over je eigenaarschap dan acceptatie. Acceptatie klinkt alsof het je overkomt en dat is precies niet hoe ik het zou willen zien." Jouw rol moet in te passen zijn naast wat er is. En ja dat kan je acceptatie noemen, maar toeslagenaffaire laat ook zien dat "computere says no" dus doe ik dat, want computers zijn slim. Die mate van acceptatie wil je niet, je wilt het kritische denken erin houden.

Veel XAI gaat over uitleg blackbox maar is dat iets waar de burger wat mee opschiet? "Je woont in een woonwagen dus daarom ben je een target" Die foutescore kaart van gemeente was letterlijk dit. Gaat meer over is het uit te leggen waarom deze specifieke criteria is meegenomen bij het maken van deze beslissing? Is dit werkelijk waar je op wilt selecteren? Het pad waarop jij er anders uit was gekomen gaat zo, en had de overheid hier iets aan kunnen doen, en is het de burger of overheid zelf die zegt, dit systeem heeft een fout. "Er zit een kant in van transparant en beter uitleggen van hoe je er tot komt. Maar het is niet meteen van alle kaders kloppen en het gaat alleen maar over het detail in het midden." Er kunnen echt wel wat grotere issues zijn die je wel ook moet meenemen.

Context

De mate van mogelijke ambiguïteit en zware impact op mensen je zeker een factor om over na te denken. Bijvoorbeeld ie parkeervergunning daar kan veel aan geautomatiseerd worden omdat veel van die gegevens goed te verifiëren zijn. En procesjes zijn waar niet heel veel menselijke controle op nodig is. Nog steeds wil je denk ik wel in alle gevallen nadenken over wat betekent het om een systeem te gamen, wat betekent het om fouten te maken? Het systeem zal niet perfect zijn. Maar je kunt idd hier meer uit handen geven.

Automatisering kan helpen om je bewust te maken van verschillen in de wereld (bijvoorbeeld oorlog, en om empathie mee te nemen). Waar ik bezorgt over ben: in de toeslagenaffaire zat dit, als je normaal een fout maakt dan is er een betalingsregeling. Maar als je van kwade opzet wordt beschuldigd dan is er geen genade. En daar ging het mis. Mensen kregen geen genade en werden als fraudeur aangemerkt. Er was een slechte incentive om snel geld terug te verdienen en de banen van de werknemers zouden er aan gaan als ze niet genoeg zouden binnenkrijgen.

Ambtenaar een mate van invloed hebben. Niet elke willekeurige ambtenaar moet elke willekeurige verbetering kunnen doen. Daar zullen toch mensen voor opgeleid moeten zijn. Maar hang er wel een soort continu onderhoudsbeeld aan. Je bent verantwoordelijk om er dynamisch in aan te passen als het daarom vraagt. En niet om een nieuwe set van instructie naar beneden te krijgen

Algoritmes gebruiken als spiegel voor de ambtenaar. Nogmaals kookrecept. Als men niet exact het recept volgt kun je daar over discussiëren als iemand bijvoorbeeld minder suiker doet maar wel goed kan koken. Die is dus expert. En ga je kijken waarom minder suiker eventueel ook kan. Die discussie wordt dan ook concreter gekaderd.

B.9 INTERVIEW 8, POLICY ADVISOR ON ETHICS

Algorithmic decision-making

Schaalvoordeel is voor een deel een groot voordeel. Het is bijna onmogelijk om alle beslissingen door mensen te laten nemen. Niet genoeg mensen voor. Er zijn ook een heleboel simpele beslissingen, die minder complex zijn en waarbij een minder lastige afweging gemaakt moet worden gemaakt.

Voor een ander deel kan je mbv algoritmes bepaalde biases uit processen halen, hangt van de implementatie af hoe je dit gaat realiseren. Wat je hier uithaalt ligt eraan hoe je het algoritme inbedt in de context van je organisatie. Vragen hierbij zijn: Hoe zijn mensen opgeleid? Wat is het kennisniveau tov algoritme? Wat voor cultuur heb je binnen je organisatie? Hoe kijkt men naar professionaliteit, technologie, etc? Zitten nog mitsen en maren aan maar kan in theorie.

Ook kan het interessant zijn om algoritmes te gebruiken om uitzondering uit te lichten, op zoek te gaan naar waar maatwerk nodig is. Waar de reguliere processen tot uitval zouden leiden, of wanneer er liever toch meer menselijke aandacht nodig is. Ook dit is een mogelijk voordeel mits goed geïmplementeerd in processen en systeemarchitectuur.

Schaalvoordeel heeft ook met efficiency van inzet van middelen te maken. Dit is (denkt de interviewee) "het algemene perspectief waarom heel veel automatisering plaatsvindt". Dit heeft ook te maken met voorspelbaarheid, regelmaat en consistentie. De schaal is daarin een breed concept, met daaronder aspecten zoals efficiency, repeteerbaarheid, betrouwbaarheid, kwaliteit, etc.

Criteria

Uiteindelijk is de impact die het kan hebben op de burger die erdoor geraakt kan worden het belangrijkste criterium. Dat is ook nu de fundamentele discussie in algoritme register, AI ACT, etc. Het gaat allemaal over hoe definieer je nu die impact, en hoe ver en hoe dicht bij de uitkomst van het algoritme gaat dat liggen. Heeft het alleen maar betrekking op het individu? Heeft het betrekking op een bepaalde populatie, die bijvoorbeeld door de inzet van het algoritme bevoordeeld of benadeeld kan worden. En breder op maatschappelijke niveau, wat doet zo'n algoritme met vertrouwen in de overheid? Dat zijn verschillende niveaus van impact waar je over moet nadenken.

Een ander aspect is de capaciteit van je systeem en van je data. Als je weet dat je data niet zo goed is, dan moet je goed nadenken over wat de rol van zo'n algoritme kan worden in een beslisproces. Een algoritme kan ook adviserend of helpend zijn. Een soort droomapplicatie zou natuurlijk zijn dat je op een intuïtieve manier kan praten tegen het algoritme en hiermee sparren, over een casus. Dit is er aan de hand, en wat moet ik doen? Kan het algoritme dan de informatie uit beleidsstukken, regels, wet halen? Zou een algoritme kunnen hanteren als een soort van sparringpartner? Hangt natuurlijk van heel veel dingen af of dit haalbaar is maar de technologie van die modellen die komt langzaam wel op dit niveau. Waarbij je constant na zal moeten denken over de impact die het heeft als je het gaat gebruiken.

Zeker bij overheidsorganisatie is menselijke controle van belang omdat je ook vanuit de Rechtsstatelijke beginselen een behoorlijke beslissing moet kunnen nemen. Hiervoor wil je toch dat er ook een mens bij de meeste gevallen mee heeft gekeken. Het liefst altijd, tenzij het zo simpel is als een adreswijziging, wat wellicht algoritmisch gebeurd. Moet je hier dan nog een mens naar laten kijken als je daar een controle op kan doen dat dat goed gaat. Zodra het gaat over beslissingen die de rechten van betrokkenen raakt, dan moet er in ieder geval een vorm van menselijke controle zijn. Het liefst heb je dat het systeem complementair is. Dus dat algoritme en mens samenwerken waarbij ze elkaar aanvullen. Voorbeeld: Processorkrachtig met consistentie van het algoritme combineren met moreelbewustzijn, creativiteit en ervaring van een beslismedewerker.

En je moet heel erg waken voor het omgekeerde, "worst of both worlds". Waarbij nuance ontbreekt, het systeem imperfect is en er iemand naast het algoritme zit die eigenlijk niet zo goed begrijpt wat hij/zij aan het doen is. Schijncontrole leidt alleen maar tot meer narigheid. Dus je moet goed nadenken over waar de complementariteit inzit, hoe kan je die realiseren, welke ontwerpkeuzes maak je in je proces om dat te faciliteren. Algoritme moet technisch goed in elkaar zitten, de datakwaliteit is heel belangrijk, maar je moet ook goed kijken naar hoe je het proces ontwerpt waarin je dat algoritme inbedt en wat je daar voor maatregelen kan nemen om daar de complementariteit te stimuleren of te waarborgen.

Dus het automatiseren kan ervoor zorgen dat ambtenaren enigszins lui worden. Of je houdt alleen maar de lastigere gevallen over welke vaak naar een team aan experts gaan. Dus er zal altijd een effect op je competenties zijn waarbij deze worden beïnvloed. Je moet hier laveren tussen computer aversie en "computere says no". Mensen moeten het willen vertrouwen maar niet blind vertrouwen waarbij ze zelf stoppen met nadenken. Daar zit een "sweet-spot". Algoritme inzetten om de complementariteiten uit te lichten en zo een betere professional te creëren.

Ook nadenken over wat je vertelt over het algoritme. Belangrijk om open te zijn over hoe dit werkt, maar tegelijkertijd moet je niet altijd met alle details uitleggen wat de score is of wat er precies gebeurd is. Dit moet wel opvraagbaar zijn wanneer dat nodig is voor een soort audittrail. Maar een bepaald niveau van onduidelijkheid kan iemand ook dwingen om zelf na te blijven denken over een situatie. En dat kan je ook combineren met dat vragen stellen van het algoritme.

Belangrijk om goed te definiëren want ander is het voor de show. "human-in-the-loop"; wat betekent dat nou? Je wilt dat iemand (de ambtenaar) daadwerkelijk invloed heeft op die "loop" en kan ingrijpen. Hierbij zul je iets moeten verzinnen om de aandacht bij de les te houden. Hierbij is dat prikkelen heel erg belangrijk. Meaningful betekent ook dat je in je proces architectuur ruimte voor betekenisvol ingrijpen van mensen in ontwerpt. Menselijke interventie wordt vaak als risico gezien waarbij dingen fout kunnen gaan. Maar dan mis je dus heel erg de betekenisvolle menselijke interactie. Wat zijn de belangrijke momenten waarop dit wel zou moeten kunnen gebeuren? Hoeft niet altijd, kun je ook over discussiëren.

Bij adreswijziging bijvoorbeeld. Als iemand belt dat deze persoon al 3 jaar verkeerd staat ingeschreven en het kan wijzigen, dan moet de ambtenaar toch handmatig kunnen ingrijpen. Ook hier zit een afweging tussen variëteit en veiligheid bij het toelaten van menselijk ingrijpen in het proces.

Persoonlijk vind de interviewee dat er voor elk algoritme een procesplaatje moet zijn over hoe het algoritme tot iets komt. En eigenlijk het liefst op N=1 niveau. Wat is er gebeurd in het algoritme. Dit kunnen dan hele ingewikkelde plaatjes zijn. Vaak ook met data mist er een groot deel over waar deze data vandaan komt, meta data. Hier moet je wel een soort van historie van kunnen zien. "Deze data is genereerd door dit algoritme, met die versie, op dit moment, etc. Dat je terug kan zien wat er gebeurd is. Uiteindelijk moet je bij een rechter toch nog kunnen uitleggen wat er gebeurd is. Of je moet een niveau van coulance hebben. Wat voor bewijslast accepteren we? Voor bijvoorbeeld herzien van een beslissing.

Context

Ook hier de impact van de beslissing. Voorbeeld parkeervergunning: gehandicapt en afhankelijk van auto maar je krijgt geen (invalide)parkeervergunning omdat het systeem jou mis kwalificeert. Als jou datakwaliteit vanaf het begin niet in orde en niet goed is, dan moet je je echt afvragen wat je hiermee gaat doen. Als je niet kan instaan voor de data die je hebt. Dan is automatiseren misschien niet ideaal of überhaupt gewenst.

Ook kijken naar "ethical life cycle" van algoritme of een digitaliseringsproduct, waarbij je moet gaan bedenken dat het niet alleen maar gaat om het ontwikkelen en implementeren maar ook om het uit faseren. Dus als jij een algoritme hebt gemaakt, en er moet na 5 jaar een ander algoritme komen. Heb jij nagedacht over hoe die transitie moet plaatsvinden, wat heb je gedaan om dit makkelijker te laten verlopen? Zit er traceability in? Dat je kan zien welke data gemaakt is met de oude en met de nieuwe versie. Die hele levenscyclus is dus ook van belang. Doet het nog wat het moet doen? Geen model drift? In de echte wereld kan iets veranderen waardoor jouw data niet meer oplevert wat het voorheen wel opleverde. Bijvoorbeeld als een variabele door bepaalde demografische factoren niet meer relevant is. Dan moet je dat ergens zien aankomen, dat je niet pas 3 jaar later merkt dat het algoritme niet meer doet wat het moet doen. Periodieke controle.

Wat gaat de levenscyclus van dit product zijn en welke maatregelen moeten er genomen worden om dit een beetje in goede banen te leiden. Updates moeten ook gedaan worden.

Werkbelasting van ambtenaren is altijd hoog maar zal niet lager worden door automatiseren. Kijk naar 40 jaar geleden. De capaciteit die vrijkomt wordt vaak opgegeten door opschaling van je processen en door meer en sneller. Dan dat ambtenaren nou meer tijd krijgen. Als je dit wil, moet je dat heel erg krachtig borgen in je managementstructuur, in je controle afspraken en in je monitoringssysteem. Dit vraagt veel.

In publieke sector word je op veel verschillende aspecten afgerekend. Gaat niet over winst. Er moet geleverd worden op responsiviteit naar de burger, rechtmatigheid, tijdigheid, politieke sensitiviteit, bestuurlijke wensen en ambities. Dus op verschillende assen moet je leveren die ook schuiven. Het is een andere organisatie dan het private domein. Ander groot verschil is dat de overheid alle burgers moet dienen en privaat een specifieke groep kiest. Hier is de laatste jaren wel al veel meer oog voor gekomen. Maar de uitdaging zit hem ook dat je nu die gesprekken met je systeemarchitecten moet hebben om het over 5-10 jaar geïmplementeerd te hebben. Dit soort dingen gaan helaas niet zo snel.

Het is beter voor de beslissing als de ambtenaar zich verantwoordelijk voor de beslissing voelt. Je wil voorkomen dat ze alleen maar doen wat er bovenin besloten is. Dus je moet ervoor zorgen dat zij die de beslissing gaan nemen het gevoel hebben dat het er ook toe doet dat ze gaan beslissen. Zonder dat ze een paralyse krijgen en niet weten wat ze moeten beslissen. Ook weer een sweet spot.

Raad van Bestuur en Minister zullen eindverantwoordelijke zijn voor hoe de dienstverlening uitgevoerd wordt. Op beslis niveau heb je wel nodig dat iemand geëquipeerd is om te begrijpen waarover deze persoon aan het beslissen is. De persoon moet niet geïntimideerd worden door de processen en de computers. Hierin moet het dus een tool voor de ambtenaar zijn. Dit vraagt wel wat van je opleidingsniveau en misschien ook wel van je cultuur. Gaat over interface naar de ambtenaar toe, inzicht geven in wat er speelt. Er zitten hier ook meerdere dimensies aan wil dit volledig slagen.

Belangrijk voor de ambtenaar om te weten, waar is mijn plek in dit proces, wat zijn de knoppen waar ik op kan drukken om dingen te beïnvloeden, en waar komt het beeld wat ik krijg vandaan? En wat de ambtenaar dan ziet, hier moet kritische op gereflecteerd kunnen worden. Dit is het ideaalbeeld. Veel beslissingen worden snel genomen ook omdat deze niet ingewikkeld zijn. Dus altijd bevraagd worden door een AI is niet nodig. Mensen willen ook verder en beweging voelen. Idealiter kan je algoritme zo'n onderscheidt ook al een beetje maken. Dit zijn de hele gladde gevallen en die kun je wellicht meer automatiseren met een soort steekproef om hier te kijken wat er uitkomt, met een gemakkelijke mogelijkheid tot herziening. In je proces kan je dan misschien met minder menselijke controle af in het stukje waar dat algoritme zit.

B.10 INTERVIEW 9, PROJECT-MANAGER

Algorithmic decision-making

Besluitvorming binnen mijn team (maar voor vrijwel alle geledingen binnen de overheid) wordt steeds gecompliceerder. Besluitvorming moet snel plaatsvinden en gelijktijdig moet er maatwerk worden geleverd. Dat bijt elkaar aangezien maatwerk toepassen tijd kost en niet zelden gecompliceerd is.

Gebruik van A.I. kan m.i. daarbij ondersteunen. Met name op het gebied van:

- efficiency: tijdsbesparing en snelheid;
- uniformiteit in besluitvorming (kwaliteit).;
- transparantie richting derden. Een beslismodel geeft een indicatie van de beoordeling.

Nu moet nog heel veel van mens op mens worden uitgezocht. Dus bellen / mailen / wie doet wat, waar, wanneer, hoelang, met welk materieel, materiaal, welke ruimte voor op- en overslag nodig. Dat kost heel veel afstemmingstijd. Als je dat kan verminderen door slimme systemen (AI) dan scheelt dat.

Als meer AI voor minder foutief geplande of uitgevoerde beslissingen kan zorgen scheelt dat natuurlijk wachttijd. Ook voor de stad en burgers veel prettiger als het aantal stremmingen afneemt, werkzaamheden sneller worden uitgevoerd en in dit domein wegen/vaarwegen minder lang gestremd zijn.

Ook betere kwaliteit, want grote kans dat AI beter kan beoordelen dan mensen (mits juiste basisinformatie beschikbaar) AI kan de vergunningverleners helpen, maar het zou ook veel breder toe te passen kunnen zijn. Projectteams / aannemers, maar ook stedelijke planners die met AI tot een veel betere volgorde van werken in de stad zouden kunnen komen. Giet alle informatie in een trechter en AI vertelt je wie, wanneer, waar kan stremmen terwijl de stad bereikbaar, leefbaar en veilig blijft. Prachtig.

Werkdruk / stress ook. Al bij het beoordelen van de vergunning, maar ook later. Als je minder vaak foutieve vergunningen hebt verleend, ook minder stress van paniekoplossingen op het moment dat blijkt dat meerdere partijen vergunningen hebben om van dezelfde ruimte gebruik te maken.

Criteria

Moeilijk. Speelt natuurlijk heel veel. AI als ondersteuning voor het maken van keuzes kan natuurlijk altijd. Dat is in dit geval niet meer dan een soort algoritme dat checkt of er geen partijen zijn die tegelijkertijd van dezelfde ruimte gebruik willen maken. Level of automation loopt dan van een beetje hulp geven aan vergunningverleners, tot aan het volledig wegcijferen van mensen die vergunningen beoordelen. Dat aanvragers alleen nog iets moeten intikken en dat meneer AI verteld of en wanneer je iets mag bouwen / stremmen.

Ik denk dat een risico / faalfactor dan voortkomt uit de verantwoordelijkheid voor de beslissing. Wie is er verantwoordelijk voor AI als het mis gaat? Wie neemt de beslissingen nu? Het waarborgen van kwaliteit blijft het de grootste rol. Gezien recente maatschappelijk ontwikkelingen (toeslagenaffaire) ligt de nadruk onverminderd groot op maatwerk/menselijke maat bij besluitvorming in het sociaal domein. Toepassing van A.I. staat daar haaks op maar is m.i. wel de oplossing voor veel obstakels.

Acceptatie van ambtenaren maar met name de maatschappij is daarbij een belangrijke factor.

Ander risico is ook dat we te afhankelijk worden van AI. Als verkeerskundigen en co allemaal niet meer de kennis hoeven te hebben over Amsterdam, de routes, de werkzaamheden enzo dan worden we afhankelijk van AI in het geven van kloppende vergunningen. Als er dan iemand is die kwaad wil en de AI kan beïnvloeden, dan kan dat gevaarlijk worden. Zo geldt ook dat het gevaarlijk kan zijn als alleen grote bedrijven of bepaalde organisaties weten wat ze moeten doen om een positief advies / vergunning van de AI te krijgen.

Context

Je kunt veel benaderen vanuit Risico = Kans * Gevolg. Gaat hier om de kans dat de AI iets verkeerd doet en de consequentie daarvan. Voor projecten delen we risico's in op: Tijd , Geld, Veiligheid , Imago, Kwaliteit, Omgeving, waarbij omgeving weer van alles kan zijn. Je kunt ook RISMAN bekijken voor invalshoeken van risico's: https://nl.wikipedia.org/wiki/RISMAN.

Maar goed, voor het interview met mij als BLVC'er heb je vast wat aan die categorieën Tijd, Geld, Veiligheid, Imago, Kwaliteit, Omgeving. Ik zou dan een inschatting willen van de kans dat AI het mis heeft en inschatting van de gevolgen op die zes factoren voordat ik besluit of en in welke mate AI voor mij mag (helpen) besluiten.

Als het om iets tijdrovends gaat, dus waar ambtenaren nu veel tijd mee kwijt zijn, en de impact van verkeerde beslissingen is klein. Prima dat AI dan alle beslissingen neemt kan ik me voorstellen. Met de notitie: beste burger, mocht deze brief niet kloppen, laat het dan even weten.

Belangrijkste factor is dan de kwaliteit van de AI in het geven van kloppende adviezen. De mate waarin je de AI kan vertrouwen. Dit heeft ook te maken met in hoeverre de ambtenaren met een AI om kunnen gaan en of zij hier voor heropgeleid moeten worden. Er komt ook een hogere acceptatie hierdoor bij de ambtenaren. Ook het ontwikkelen van nieuwe skills bij het werken met ai, stukje training komt hierbij kijken. IT integratie en kennis van implementatie is ook hoog nodig. Zeker omdat heel veel overheden en gemeentes vastzitten aan een bepaalde supplier. En daar zijn regels over. Dus zeker belangrijk om te vertellen wat kan wel en wat kan niet op basis van het contract wat je hebt. En met de software die je gebruikt.

Maar als het gaat om belangrijke besluiten met veel impact dan wil je vooralsnog altijd menselijk besluit. Maar goed, dit is een hele lastige. Ik heb ook weleens gelezen dat er van die psychologische onderzoeken zijn dat rechters flink beïnvloed worden door omgevingsfactoren in de mate waarin ze hoge straffen geven. In zo'n geval kan je zeggen dat een objectieve AI als adviseur erg prettig kan zijn. Zo zullen er voor ambtenaren vast ook dingen zijn waarbij de AI de taak eigenlijk beter/wenselijker uitvoert dan de ambtenaar.

De rol van AI is afhankelijk van de impact van het besluit. In het sociaal domein is de impact van een besluit vaak groot. Denk bijv. aan de gevoeligheid binnen de afdeling jeugd. Die impact is kleiner bij het al dan niet verstrekken van een parkeervergunning. Hoe groter de impact hoe belangrijker de rol van de ambtenaar in de besluitvorming. Drukte is daarbij van minder groot belang.

Voor op de korte termijn zou ik in mijn situatie wel de check willen doen op het advies dat de AI mij geeft over het wel/ niet vergunnen van ruimtegebruik in de stad. Belangrijk begrip is dan nog navolgbaarheid. Als de AI mij precies kan uitleggen waarom het op een bepaald advies is gekomen, dan kan ik die stappen checken, maar als de AI alleen JA, vergund of NEE, afgewezen zegt, dan durf ik 'm nog niet zo snel te vertrouwen.

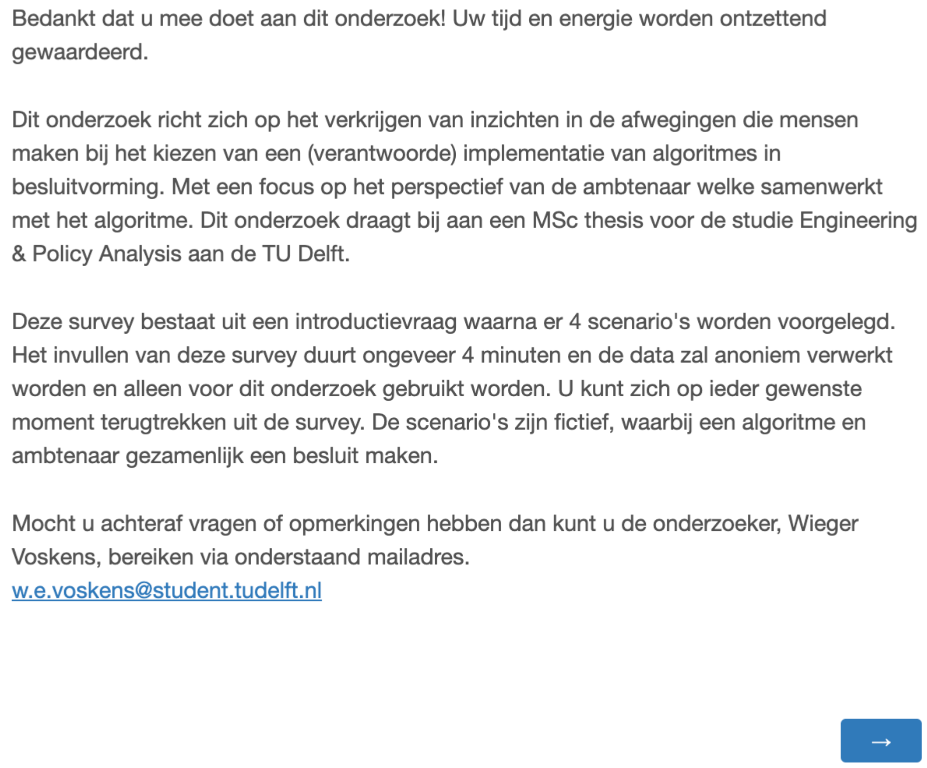
Verantwoordelijkheid is ook heel moeilijk. Beetje zoals de automatisch rijdende auto. Wie is dan verantwoordelijk voor een ongeluk. Je zou nu in ieder geval degene die de AI (gedeeltelijk) gebruikt verantwoordelijk stellen, zoals je dat nu ook doet voor bestuurders die gedeeltelijk gebruik maken van automatisch rijden, cruise controll etc.

De ambtenaar neemt het besluit en draagt hier van de verantwoordelijkheid. A.i. werkt daarbij sturend en ondersteunend. Een (verkeerd) besluit blijft altijd onder verantwoordelijkheid van de afdeling die het besluit neemt. Bij het nemen van een besluit zijn er nu waarborgen ingeregeld (2 paar ogen principe). Dat blijft in stand.

C APPENDIX C:

C.1 VIGNETTE EXPERIMENT

For the vignette experiment, the following four scenarios are present for 2 different cases. These are mentioned in chapter 5, but are stated here as well together with the other, open and demographic questions that are asked in the survey experiment. Before presenting these cases and scenarios the opening statement and the introduction question are stated below.



Bedankt dat u mee doet aan dit onderzoek! Uw tijd en energie worden ontzettend gewaardeerd.

Dit onderzoek richt zich op het verkrijgen van inzichten in de afwegingen die mensen maken bij het kiezen van een (verantwoorde) implementatie van algoritmes in besluitvorming. Met een focus op het perspectief van de ambtenaar welke samenwerkt met het algoritme. Dit onderzoek draagt bij aan een MSc thesis voor de studie Engineering & Policy Analysis aan de TU Delft.

Deze survey bestaat uit een introductievraag waarna er 4 scenario's worden voorgelegd. Het invullen van deze survey duurt ongeveer 4 minuten en de data zal anoniem verwerkt worden en alleen voor dit onderzoek gebruikt worden. U kunt zich op ieder gewenste moment terugtrekken uit de survey. De scenario's zijn fictief, waarbij een algoritme en ambtenaar gezamenlijk een besluit maken.

Mocht u achteraf vragen of opmerkingen hebben dan kunt u de onderzoeker, Wieger Voskens, bereiken via onderstaand mailadres.
w.e.voskens@student.tudelft.nl

Figure C.1: Introduction to the survey (*in Dutch*)

Before presenting the scenarios and the cases the respondents are asked through a multiple choice question to mention the domain where they are working in. The options are the following:

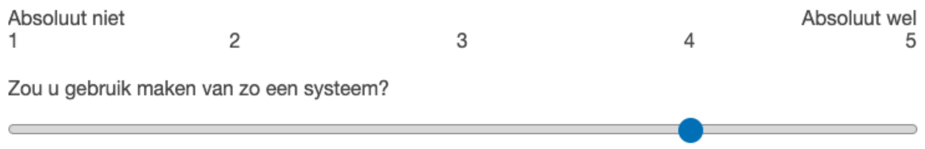
1. Working in the public domain
2. Working in the scientific domain
3. Working (with AI) in the private domain
4. Other
   a) *With the option to mention in what other domain they are working*

After this option, every respondent is given the opportunity to disclose more about their role or function. This option is voluntary.

Table C.1: Case descriptions as provided in the vignette experiment (in Dutch)

Casus 1: Bijzondere Bijstand
Ambtenaren beslissen, op basis van een aantal criteria, of een burger recht heeft op de zogeheten Bijzondere Bijstand (link naar Rijksoverheid). Deze beslissing hangt af van gevoelige data zoals de financiële of gezinstoestand van een individu. Hieronder worden tweescenario's geschetst waarin een algoritme samen met de ambtenaar deze beslissing neemt. Dit is een hybride vorm waarin de mate van controle verschilt voor de ambtenaar en algoritme voor de verschillende scenario's.
English
Civil servants decide, based on a number of criteria, whether a citizen is entitled to the so-called "Bijzondere Bijstand" (special welfare) [Rijksoverheid, 2023b]. This decision depends on sensitive data for example the financial or family situation of an individual. Two scenarios will be described below in which an algorithm has been implemented in the decision-making process where it supports the civil servant in coming to this decision. This is a hybrid form in which the level of control differs for the algorithm and for civil servants in different scenarios.
Casus 2: Parkeervergunning .
Ambtenaren beslissen, op basis van een aantal criteria, of een burger recht heeft op een parkeervergunning (link naar Rijksoverheid). Deze beslissing hangt af van een aantal voorwaarden zoals een kenteken dat op naam van de betreffende burger moet staan. Dit kan per gemeente verschillen. Voor nu gaan we ervan uit dat het niet gevoelige data is welke gebruikt wordt (woont de burger in dit gebied en is het inderdaad de (enige) auto van de burger). Hieronder worden twee scenario's geschetst waarin een algoritme samen met de ambtenaar deze beslissing neemt. Dit is een hybride vorm waarin de mate van controle verschilt voor de ambtenaar en algoritme voor de verschillende scenario's.
English
Civil servants decide, based on a number of criteria, whether a citizen is entitled to a parking permit [Rijksoverheid, 2023a]. This decision depends on a number of criteria such as a licence plate that has to be registered to the name of the corresponding citizen. Some criteria may differ slightly per municipality. For now, it is assumed that no sensitive data will be used (the citizen should be living in the corresponding area and this car is indeed their only car). Two scenarios will be described below in which an algorithm has been implemented in the decision-making process where it supports the civil servant in coming to this decision. This is a hybrid form in which the level of control differs for the algorithm and for civil servants in different scenarios.

For each of these vignettes, the respondent is asked to rate the likeliness of using this scenario from the perception of the civil servant using a slider as presented below.



Zou u gebruik maken van zo een systeem?

Figure C.2: Example of a slider in the survey (*in Dutch*)

Table C.2: scenario descriptions as provided in the vignette experiment (*translated*)

Case 1: Special Welfare
Scenario 1
By making use of the algorithm the decision can be taken within 5 minutes, which normally takes about an hour. This enables civil servants to spend more time in direct contact with citizens having more time to explain certain decisions that have been made (more attention for the human). The algorithm here advises a certain decision to the civil servant. There is little control for the civil servant as they base their choice on the advice given by the algorithm. Next to this, it is unclear how the algorithm got to this advice and how certain criteria have been traded off by the algorithm. The result is that the civil servant has little impact on the making of the decision.
Scenario 2
By making use of the algorithm, the data, needed to come to a decision, is directly analysed and presented in an overview for the civil servant. This results in the civil servant being able to take a well-informed decision within 45 minutes instead of 1 hour, which it normally takes. The civil servant thus has more time to explain the decision to the citizens. How the algorithm analyses the data is comprehensible and accessible for the civil servant. The civil servant here can intervene and have another look at the data when they have doubts about the overview provided by the algorithm.

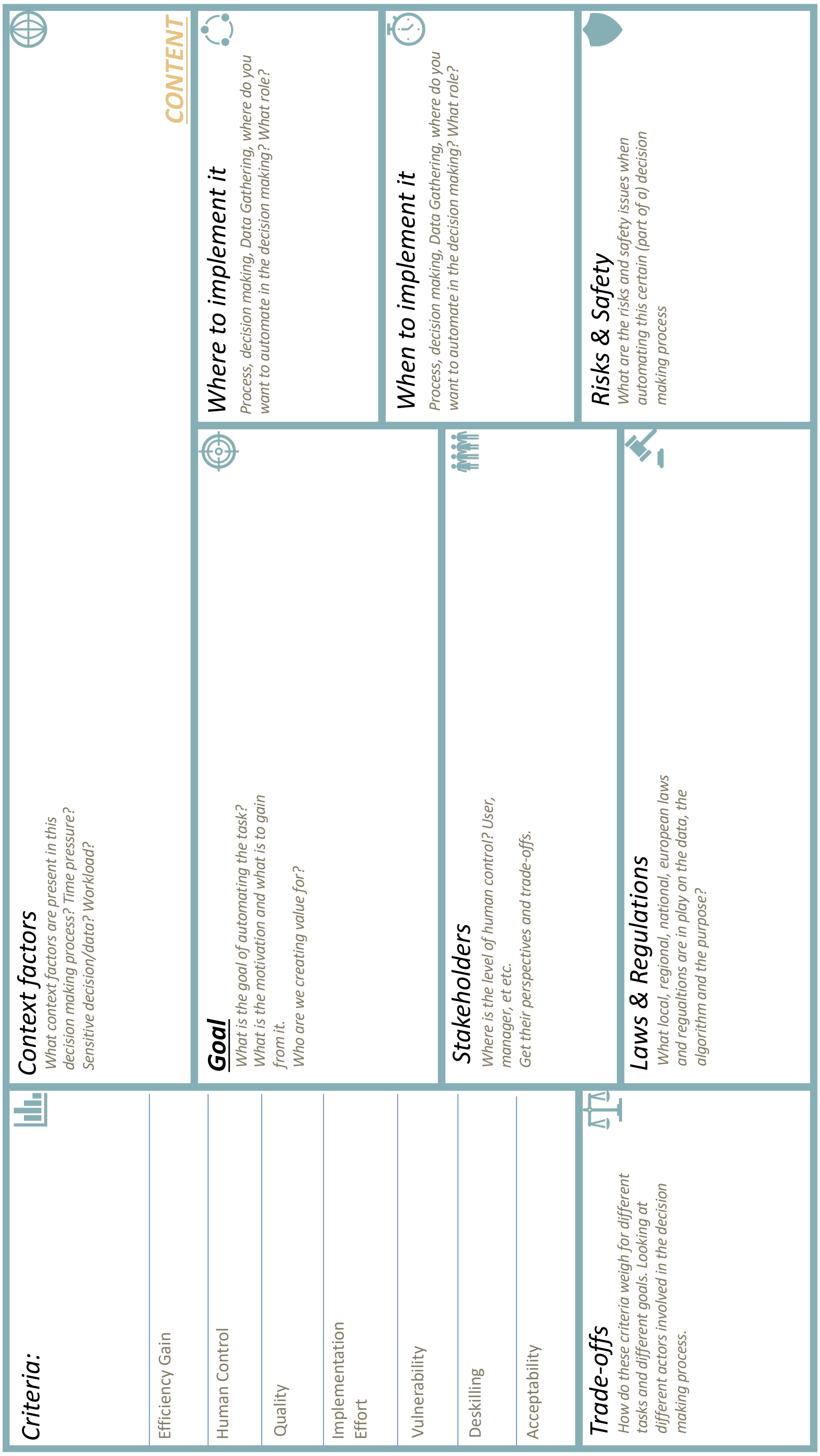
Case 2: Parking permit
Scenario 3
By making use of the algorithm the civil servant can grant or decline a parking permit within 5 minutes. This enables civil servants to spend more time on personal attention with more complex decisions. The algorithm makes a decision on issuing a parking permit based on several criteria and the civil servant's only task is to implement this decision or not. How the algorithm comes to a decision is not known but if there appears a case that deviates from what normally happens, the civil servant can intervene.
Scenario 4
By making use of the algorithm the civil servant can grant or decline a parking permit within 15 minutes. This enables the civil servant to spend more time providing an explanation for this decision. The algorithm gives a clear overview of the criteria and provides the civil servant with advice. The civil servant can deviate from this advice when this is considered not the right decision in the eyes of the civil servant. How the algorithm comes to this advice is unknown.

D APPENDIX D:

D.1 CANVAS

The canvas as presented in this appendix is suggested as a tool to conduct further research on the remaining aspects that aid in determining a responsible level of automation. It is also thought that such a canvas can help in guiding policymakers willing to experiment with the implementation of algorithms and AI in the public domain. The research focuses on the criteria, trade-offs and contextual factors parts of the canvas, though it touches upon the others and also suggests some guidelines and requirements for aspects such as the role of the civil servant, lacking knowledge from policymakers, responsibility aspects, European and national guidelines, and risk and safety assessments. The canvas is presented to show how this research could be further extended and how a concrete tool could be established that is founded on scientific concepts and evidence, aiming to advise policymakers but also AI and algorithm developers in the private domain, where they can already see what they need to comply to, in order to be even considered by public organisations.

The canvas consists of two sides where the first one explains what needs to be filled in in all of the blocks and the other side elaborates briefly on some possible methods and techniques that could be used to give answers to these questions and to fill in the blocks. This canvas is a concept version and plenty of adjustments can and need to be made for it to be implementable. These adjustments could also differ per domain or per decision-making process that it is used for. It is encouraged to work on it, improve it, change it, test it, and experiment with it.



Criteria:

Efficiency Gain

Human Control

Quality

Implementation Effort

Vulnerability

Deskilling

Acceptability

Context factors

What context factors are present in this decision making process? Time pressure? Sensitive decision/data? Workload?

CONTENT

Goal

What is the goal of automating the task?
What is the motivation and what is to gain from it.
Who are we creating value for?

Where to implement it

Process, decision making, Data Gathering, where do you want to automate in the decision making? What role?

Stakeholders

Where is the level of human control? User, manager, et etc.
Get their perspectives and trade-offs.

When to implement it

Process, decision making, Data Gathering, where do you want to automate in the decision making? What role?

Trade-offs

How do these criteria weigh for different tasks and different goals. Looking at different actors involved in the decision making process.

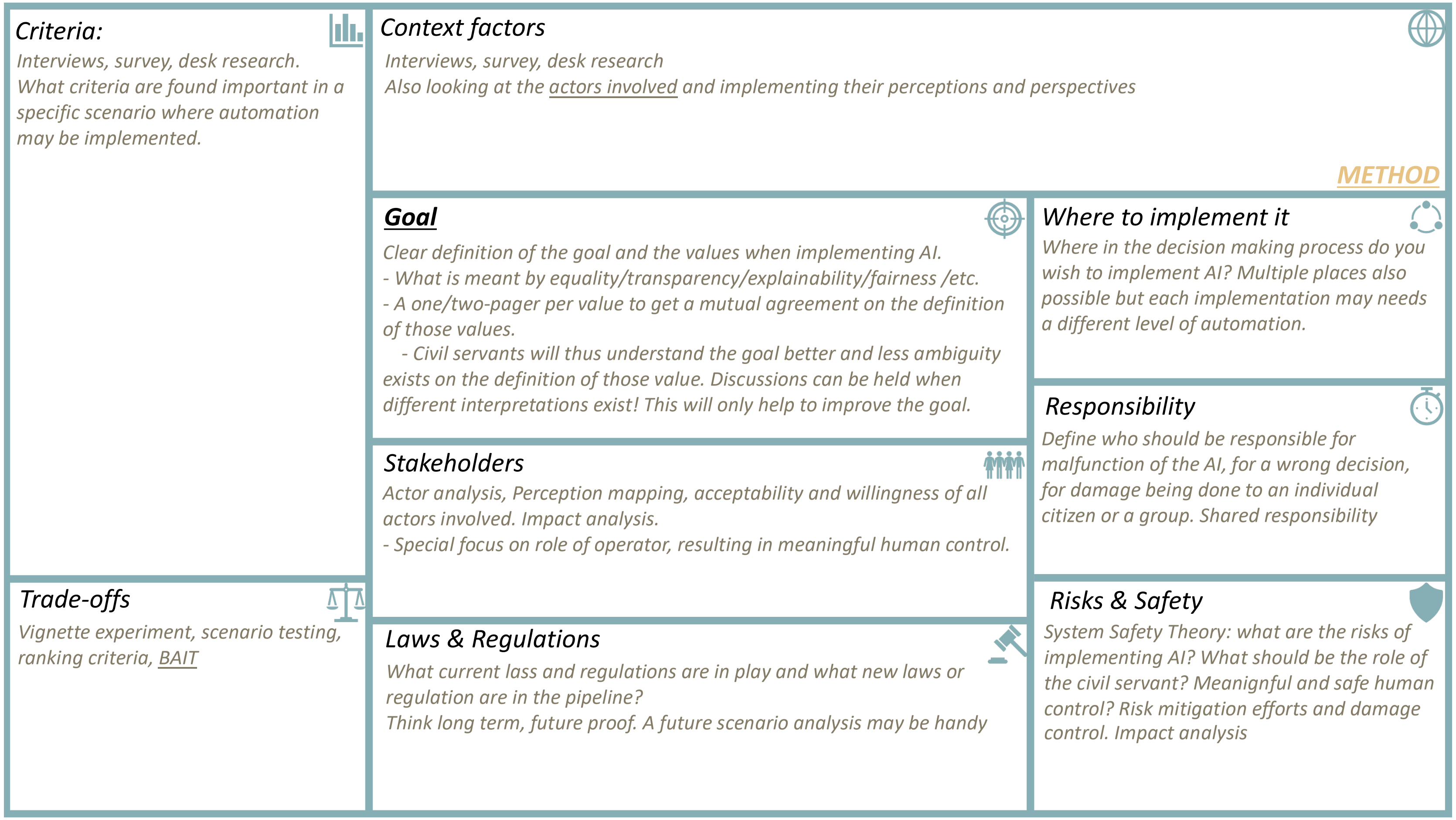
Laws & Regulations

What local, regional, national, european laws and regualtions are in play on the data, the algorithm and the purpose?

Risks & Safety

What are the risks and safety issues when automating this certain (part of a) decision making process

Figure D.1: Canvas



Criteria:

Interviews, survey, desk research.
What criteria are found important in a specific scenario where automation may be implemented.

Context factors

Interviews, survey, desk research
Also looking at the actors involved and implementing their perceptions and perspectives

METHOD

Goal

Clear definition of the goal and the values when implementing AI.
- What is meant by equality/transparency/explainability/fairness /etc.
- A one/two-pager per value to get a mutual agreement on the definition of those values.
- Civil servants will thus understand the goal better and less ambiguity exists on the definition of those value. Discussions can be held when different interpretations exist! This will only help to improve the goal.

Where to implement it

Where in the decision making process do you wish to implement AI? Multiple places also possible but each implementation may needs a different level of automation.

Responsibility

Define who should be responsible for malfunction of the AI, for a wrong decision, for damage being done to an individual citizen or a group. Shared responsibility

Stakeholders

Actor analysis, Perception mapping, acceptability and willingness of all actors involved. Impact analysis.
- Special focus on role of operator, resulting in meaningful human control.

Trade-offs

Vignette experiment, scenario testing, ranking criteria, BAIT

Laws & Regulations

What current lass and regulations are in play and what new laws or regulation are in the pipeline?
Think long term, future proof. A future scenario analysis may be handy

Risks & Safety

System Safety Theory: what are the risks of implementing AI? What should be the role of the civil servant? Meaignful and safe human control? Risk mitigation efforts and damage control. Impact analysis

Figure D.2: Canvas, methods and techniques

COLOPHON

This MSc thesis: *To AI or not to AI?* How do civil servants trade-off different levels of automation in local governments? was typeset in LaTeX, using the arsclassica package by Lorenzo Pantieri. Cover figure: AI-generated via Microsoft Bing. *Prompt*: ”To AI, or not to AI”, Shakespeare orating in the 21st century on his new play on AI, in the style of Salvador Dali