A User-Centered Explainable AI Visualization Study for Enhancing Decision Making in Law Enforcement

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Human Computer Interaction

Author:	Kim van Genderen
Student ID:	6497039
Supervisor:	Dr. E. Dimara
Second supervisor:	Prof. dr. F.J. Bex
External supervisor:	Marcel Robeer MSc
Date:	November 1, 2023
Location:	Utrecht

Universiteit
Utrecht

Acknowledgements

This thesis entitled "A User-Centered Explainable AI Visualisation Study for Enhancing Decision Making in Law Enforcement" serves as the final project for my Master's degree in Human Computer Interaction at Utrecht University. It marks the end of my academic career and I am grateful for the invaluable knowledge and enjoyable experiences gained during my studies.

I conducted research on a very interesting and relevant topic, which is related to both my Bachelor's degree in Artificial Intelligence and my Master's degree in Human Computer Interaction. I was therefore pleased to be able to put my knowledge from both studies into practice during a research internship at the National Police Lab AI. This master's thesis is the final product of this research internship.

This study would not have been possible without a number of people. First of all, I would like to express my deepest appreciation to my research supervisor, Evanthia Dimara, for her invaluable guidance throughout my master's thesis. Her eagerness to support my research, perceptive insights, extensive scientific knowledge in the realm of data visualisation, and constructive feedback have greatly added to the calibre of my research. Secondly, I would like to thank my second reviewer, Floris Bex, for providing me with valuable feedback on my research proposal, which has enabled me to improve my research. Additionally, I would like to express my appreciation to the National Policelab Artificial Intelligence for providing me with the opportunity and resources to conduct my master's thesis within this interesting organisation. In particular, I would like to thank Marcel Robeer for guiding me through the organisation. His extensive knowledge, never-ending guidance and support, and fair and constructive criticism have enabled me to gain a deeper understanding of both interpretable Machine Learning and the research process.

Last but not least, I would like to thank my parents, friends, and family for their unwavering support and encouragement during my years of study and throughout my research. Without them, this accomplishment would not have been possible.

Kim van Genderen

Abstract

Police need to be able to analyse large amounts of data in order to enforce the law. Nevertheless, this task cannot be solely completed by humans, and therefore requires the utilization of Machine Learning (ML) models. However, a concern arises in terms of the lack of transparency of these models, which could have major consequences in high-stake scenarios. Hence, it is critical to provide explanations if we anticipate using such models in fields such as law enforcement. However, many visualisations have been developed to explicate Machine Learning (ML) models for data scientists instead of decision makers. This creates an issue as data scientists have distinct objectives when interacting with an ML model compared to decision makers. While data scientists possess technical knowledge, they lack domain knowledge and seek solutions to improve the model. Conversely, decision makers utilise the model's output to inform their decision making processes. They possess domain knowledge but lack technical expertise. Due to distinct characteristics and different requirements when dealing with ML models, they also require a different explanations. In collaboration with the National Police Lab Artificial Intelligence (NPAI), this research developed a way to effectively visualise local explanations of ML models for decision makers in the law enforcement domain. We focused on decision makers within public order and safety domains. The interviews unveiled several prerequisites that were integrated into the design. The evaluation demonstrated that decision makers comprehended the visualization and that the tool facilitated decision making. Nevertheless, it emerged that the explanation was not entirely comprehensible to the decision makers. They could pinpoint the characteristics that influenced the classification of the risk and identify the risk that the model attributed to the incident. However, they lacked the ability to discern which features made a larger contribution, and the uncertainty score proved challenging to interpret.

Contents

Acronyms

Notation	Description
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
iML	interpretable Machine Learning
LEAs	Law Enforcement Agencies
ML	Machine Learning
XAI	Explainable Artificial Intelligence

1. Introduction

Law enforcement is the activity of enforcing the law by detecting, punishing, deterring, or rehabilitating people who violate the norms and rules of society. This activity is conducted by Law Enforcement Agencies (LEAs), who are responsible for maintaining the rule of law, ensuring compliance with the law, and providing assistance to people in need [1, 2, 3]. In other words, LEAs ensure that citizens comply with the law by acting against crimes and violations, and provide assistance to support different help organisations (e.g., ambulances and fire departments).

Each country has their own unique law enforcement structure [4]. However, in most countries, LEAs operate on three levels: municipal, regional, and national. Nevertheless, there are places with more or fewer levels. For example, in parts of Canada there is only a national police force [2], while the Netherlands also has a department responsible for airports [3]. The level determines the rights of the LEAs. For example, LEAs operating at the municipal level are only entitled to enforce the law within their municipality.

The most well-known LEA is the police, but international bodies such as Europol and Interpol also fall under LEAs. The police have a special role in society, as they have the constitutional right to use force to enforce the law. The police can only do use force when strictly necessary and when it is required to carry out their duty [5, 6, 7]. Europol is responsible at European level for the organisation and co-ordination of cross-border operations against dangerous criminal groups within the EU. However, Europol does not have the right to use force, but supports the work of the police and justice systems in the EU Member States [6]. Interpol, on the other hand, is a global organisation, which does not have the right to use force, but is also a provider of information to countries [8].

In order to enforce the law, the police have to analyse many different sources of information. Examples include, the analyses of recorded crimes, noting the location, personal details and findings in order to identify patterns, trends and correlations [5]. In addition, surveillance cameras or traffic cameras are analysed to determine whether someone drives through a red light or at a high speed [9]. Forensic analysis of computer hard drives, laptops or mobile phones may also be required as part of a criminal investigation [10, 5]. Furthermore, police utilize social media to improve resource allocation [9, 11].

The police thus use very large data sets that grow exponentially over time, known as big data [12, 13]. However, the sources used by police are often unstructured. Unstructured data is not organised in a predefined manner, making it difficult to process and analyse. Examples include blog posts, PDF files, videos, web pages and numerous more [14]. Structured data, on the other hand, is organised in a well-defined manner, such as in tabular form, and is therefore easier to analyse [14].

The aim of the police is to analyse this big data to discover hidden patterns and insights in order to enforce the law [15]. Nevertheless, analysing big data is time consuming and very complex as the datasets are unstructured and growing in size every day. However, the analysis can inform us of new patterns and trends that would probably go unnoticed if we only looked at each dataset individually [12]. Therefore, there is a need to automate the analysis, and ML models are often used to do this, since ML models can help guide important and complex human decisions [16] by finding trends and patterns in big data [17].

ML models can benefit us in many ways: they can entertain us, by providing suggestions for films, TV shows and music based on the user's preferences and previous interactions; help us save lives and cope with disasters; and can relieve us from boring, hard, or dangerous work [5]. The models are already widely used in various domains [16], since they are well suited to manage complex, data-intensive tasks [5], such as diagnosing diseases, checking credit card systems for fraudulent behavior, detecting cybersecurity threats, and predicting the risk of committing future crimes [5, 18, 19, 20].

ML models also potentially have beneficial applications within law enforcement, such as user-friendly services for citizens (e.g., interactive forms or chatbots to file reports), processing large amounts of data (e.g., from police reports or from seized digital devices), automated surveillance, finding case-relevant information to support an investigation and prosecution, predictive policing, and in improving productivity and paperless workflows [5].

However, the demand for transparent ML models is growing, especially when the models are used for tasks that have high social impacts, as in the law enforcement domain [19, 5, 21, 22]. The reason for this demand is that decisions in high-stake environments can have major consequences on individuals and society as a whole [5, 23]. In addition, algorithm transparency is also important to hold organisations responsible and accountable for the development and use of ML models [5].

1.1 Problem Statement

According to the research conducted by Bertini and Lalanne [24], visualisations can aid ML model interpretation. Spinner et al. [22], even say that visualisations are a natural way to obtain interpretable explanations for humans. However, current research focuses on designing interfaces to visualise the behaviour of ML models for data scientists rather than decision makers [25, 26, 27]. This is a problem because data scientists have different goals in mind when interacting with a model than decision makers [26, 28].

Data scientists are interested in understanding, refining and diagnosing models to improve them [29]. They are familiar with the use and development of ML models and therefore have technical knowledge [16, 22]. To gain insight into how the model works, the data scientists use statistics such as accuracy, precision and F1 score [30, 22]. Decision makers, on the other hand, are non-experts in ML, but are experts in the domain [16]. The aim of this group is therefore not to improve the models, but rather to use the insights from the model to guide their decision making.

Currently, the decision makers in various industries depend on data scientists to analyse complex data, extract insights, and provide recommendations that inform their decision making processes [26, 28, 31]. The data scientists communicate the outputs and explanations of the ML models with the decision makers [31]. This means that the decision makers are forced to use the kind of information that the data scientists think is relevant and useful. However, the data scientists have different goals in mind compared to the decision makers [26, 28]. As a consequence, the information provided by the data scientists might not be the information the decision makers need to make their decisions. In addition, the different levels of background knowledge also imply different needs and requirements regarding the explainability of the models [22].

We need to ensure that visualisations are designed to meet the needs and requirements of decision makers if we are to provide visualisations of explanations of ML models for law enforcement decision makers. This is certainly important in law enforcement, where decision makers make decisions that will significantly affect people and society.

Generally, one can distinguish two types of explanations: global and local explanations [32, 19]. Global explanations describe how the overall ML model works. In contrast, local explanations provide behavioral descriptions of specific input-output pairs. In other words, they describe how the ML model behaves for an individual prediction. In this research, the emphasis will be on local explanations, as we are interested in how the explanation of a model can contribute to decision making processes in law enforcement. Our focus here has been on specific input-output pairs decisions, such as recognising the number plate of a vehicle that is exceeding the speed limit.

The current challenge in Explainable Artificial Intelligence (XAI) research is the lack of high-quality user evaluations [33, 34, 35, 36, 18]. This issue arises from incomplete or altogether absent evaluations in some studies, while others provide very limited detail on how and what has been evaluated [33, 34, 18]. This research aims to address this limitation by using the nine-stage framework proposed by Sedlmair et al. [37], which provides a practical guidance on conducting a design study and working with domain experts. The framework allows for a comprehensive description of the steps that are involved in the design and evaluation of visualisation tools, as well as all of the intermediate stages. Further explanation of the framework will be provided in Section 1.5.

1.2 Research Questions

The aim of this research is twofold: (i) to create a design language that explains the local decisions of an ML model to decision makers within the law enforcement domain, and; (ii) to use this design language to create a better understanding of the explanations so that they can be incorporated in the decision making process of the decision makers.

We have formulated the following research question to address our main aim:

RQ. How can local explanations of ML models be effectively visualised to enable law enforcement decision makers to better understand and incorporate them into their decision making process?

In order the answer this research question, we formulated the following sub-questions:

SQ1. What types of explanations are common in the field of XAI, especially relevant to law enforcement applications?

We examine existing methods of explaining ML models to gain an understanding of how these work. Furthermore, we look at how these methods are being used in existing interfaces to gain an understanding of the domains and types of users for whom this is being done. We then use these findings to develop our interface.

1.1 What methods are there to make ML models interpretable, particularly for law enforcement?

1.2 How are these methods visualised in existing tools relevant to law enforcement?

SQ2. Who are the stakeholders that potentially interact with ML models in the law enforcement domain, and what are their needs related to the decision making process?

By means of semi-structured interviews, we want to discover the tasks performed by decision makers in the law enforcement domain. We want to recognise what the needs of the decision makers are when making decisions, what data they use, with whom and how they communicate this information, and whether they use any tools to help them make their decisions. The findings from the interviews are used to design an interface to support decision makers in their decision making task.

2.1 Who are the stakeholders that can potentially interact with an ML model in the law enforcement domain?

2.2 What types of decisions must be made in the law enforcement domain when utilising ML models?

2.3 What types of data do stakeholders in the law enforcement domain use when interacting with ML models?

2.4 What technology and tools do these stakeholders use to guide them in their decision making process?

2.5 How do these stakeholders communicate their information with others? This includes the methods they use for communication and the roles of the people they communicate the information with.

SQ3. How can the insights from SQ1 and SQ2 be utilised to develop an effective and interpretable visualisation of local explanations for decision makers in the law enforcement domain?

In this sub-question, we aim to synthesise the findings from SQ1 and SQ2 to create a visualisation design that presents local explanations of an ML model in an interpretable way for decision makers involved in law enforcement.

3.1 Which explanation types from SQ1 are applicable and relevant to the specific needs, requirements and context of decision makers in the law enforcement domain that we identified in SQ2?

3.2 How can we incorporates these applicable explanation types, addressing the most critical needs of the decision makers and other stakeholders in the law enforcement domain into a visualisation prototype design?

3.3 How effective and interpretable is the proposed design? To address this question, we conduct evaluations to gather feedback from target users, such as decision makers and other law enforcement stakeholders.

1.3 Contributions

This research makes both scientific and social contributions. The *scientific contributions* are as follows:

We propose an effective method of explaining the local explanations of an ML model to decision makers in law enforcement, thereby enhancing their decision making processes. Our target audience is the law enforcement domain in the field of public order and safety. The findings of the research can be used to explain the local decisions of ML models to decision makers from different fields within law enforcement.

Furthermore, the research aims in creating a shared understanding of the domain knowledge used by law enforcement decision makers, enabling them to make informed decisions relating to public order and safety. The method used to gain this shared understanding can provide a basis for gaining a shared understanding of the domain knowledge of decision makers in other domains (both within and outside law enforcement).

Finally, this research provides a detailed description of how the user studies were conducted. This contributes to the current lack of high quality user studies in the field of XAI. Our research entailed eleven interviews and three evaluation sessions.

The *societal contributions* of this research are:

The designs employed in this research to visualise local explanations of ML models can act as a foundation to visualise local explanations of ML models for other decision making processes within the law enforcement domain.

Furthermore, this research offers a guide for other organisations and companies on how to conduct a design study in their organisation or company so that decision makers can be supported in their decision making tasks by ML models. The guide suggests ways in which user studies can be conducted to identify the needs and requirements of decision makers.

1.4 Thesis Outline

The remainder of the research is structured as follows. Section 1.5 outlines the nine-stage framework proposed by Sedlmair et al. [37]. This framework serves as a guide for this paper to visualise local explanations of an ML model. Each phase of the framework will be explained and its relevance to this particular research will be elaborated upon. The choice of this framework and the phases of the framework will be the subject of discussion. Then, for each phase, we explain its relevance to this particular research.

The remaining chapters are divided into five parts, as depicted in Figure 1.1. Chapter 2 discusses the three different research areas (law enforcement, interpretable Machine Learning (iML), and data visualisation) that are central to this research. In section 2.1 we explicate the Dutch police, as this research is conducted in association with the National Police Lab Artificial Intelligence (NPAI). In section 2.2 we explore sub-question SQ1.1 by examining the distinct kinds of ML models and how we ensure their transparency to the human user. In addition, in section 2.3, we discuss how to design explanations that meet the needs and requirements of the user. Finally, in section 2.4 we will discuss why visualisations are needed in explaining ML models and which aspects are relevant to do so.

Chapter 3 deals with sub-question SQ 1.2. The aim of this chapter is to establish design themes suitable for our interface. Initially, this chapter examines the two most typically utilised methods for elucidating ML models, in section 3.1. In addition, we give a concise explanation of how these two methods work and show how they are visualised by the developers of the methods. Next, in section 3.2 we review related work on existing interfaces that use these two methods to explain ML models. The overview of existing interfaces outlines the domains and target groups for which they have been developed and how the explanations are visualised.

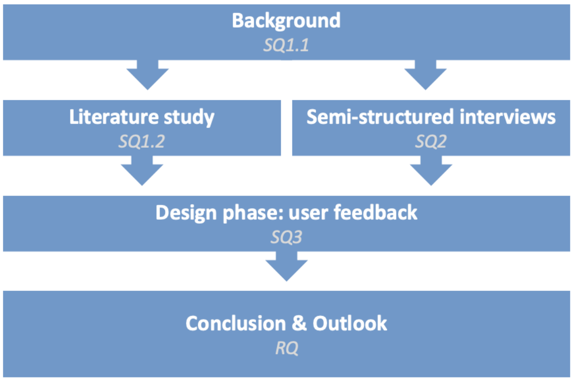


Figure 1.1: Thesis overview

Chapters 4 and 5 cover SQ2 by describing the process of the semi-structured interviews and analysing the resulting findings and insights.

Chapter 6 presents our methodology for designing the interface, drawing on the results of SQ1 and SQ2. In addition, in this chapter we will discuss the process of the feedback session and how we used the insights from each of the sessions to develop a new design. Finally, this chapter provides a detailed description of the final design.

In chapters 7 and 8, the evaluation process and the derived findings and insights are discussed. These results provide valuable input towards addressing the main research question.

Finally, the thesis is concluded in chapters 9 and 10. Chapter 9 gives a summary of the study's limitations. Chapter 10 concludes this research by providing concise answers to sub-questions SQ1, SQ2 and SQ3, and then use these findings to answer the main research question. Furthermore, potential future research directions are presented in this chapter.

1.5 Research Approach: Design Study

This study uses the nine-stage framework proposed by Sedlmair et al. [37], which is a design study method that is a practical guide on how to conduct a design study and how to collaborate with domain experts. A design study is a form of the broader concept of problem-driven research. The goal of this type of research is to design visualisations that solve real-world problems of real users [37].

Rather than just focusing on the evaluation of visualisation systems, this framework provides practical guidance from the beginning of the design of a visualisation system to its evaluation and all the steps in between [37]. This is why we use this framework in this thesis, because it is not only important to validate whether the proposed visualisation actually solves the problem in the real world, but also to identify the user's needs from the beginning, so that this information can be used in the design phase.

The nine-stage framework consists of nine different stages (*learn, winnow, cast, discover, design, implement, deploy, reflect, and write*), each divided into three different categories (*precondition phase, core phase, and analysis phase*), see

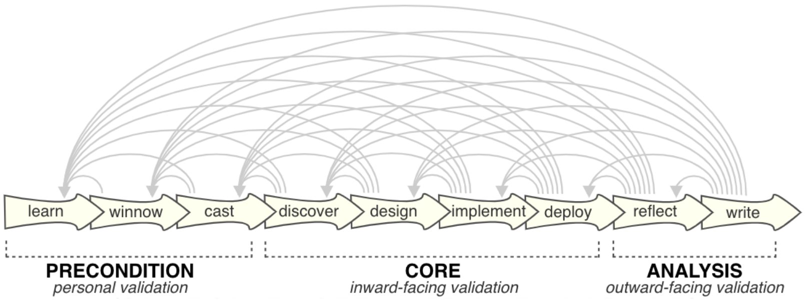


Figure 1.2: *Nine-stage framework proposed by Sedlmair et al. [37].*

Figure 1.2. Validation is important in each stage of this framework because the outputs of one stage are the inputs of the next stage. Therefore, if a wrong decision is made in an early stage, it will affect other stages because the problem will not be solved. However each stage has its own appropriate validation method, hence the three different categories [37]. We will briefly discuss each stage of this framework and how we will apply it in our research.

The validation is individual in the pre-condition phase, since it depends on the preparation of the researcher. In the core phase the validation is inward-facing, meaning that it is dependent on the evaluating findings with the domain experts. The validation in the analysis phase is outward-facing, which means that it is dependent on justifying the results of the design study.

1.5.1 Precondition Phase

The aim of this phase is preparing the visualisation researcher, and to establish and define useful collaborations with domain experts. The precondition phase consist of three different stages:

- *Learn*:
  In this stage, acquiring knowledge of the visualisation literature is central. This includes knowledge about interaction techniques, visual encodings, and design guidelines. This knowledge forms the basis for the later stages. The results of this stage will be covered in 2.4.

- *Winnow*:
  At this stage the goal is to identify the most promising collaborations. This thesis is a collaboration with the National Policelab AI (NPAI). All participants and domain experts have been recruited from the Dutch National Police. A concise overview of the organisation of the Dutch police will be covered in section 2.1.

- *Cast*:
  At this stage the different roles in the project are defined. The different roles that will be fulfilled in this research are: researchers, data scientists, decision makers, and peer students.

  - The researchers consist of myself (the student researcher), supported by my supervisors from Utrecht University and NPAI. The student researcher is responsible for conducting the study, doing the interviews, developing a design language, and evaluating the proposed idea.
  - The data scientists are the people who develop and evaluate ML models for the Dutch National police. They are therefore also a user of the model.
  - The decision makers are the people who have to make a decision with the output of an ML model. This person could also be a data scientist, but in most cases this will be another person. In this research, the roles of decision maker and data scientist are fulfilled by different people.
  - The peer students are used to conduct the pilot studies for the interviews and the evaluation sessions.

1.5.2 Core Phase

This phase consist of four stages:

- *Discover*:
  At this stage, the goal is to discover what the needs, problems and requirements of the domain experts are, to determine if and how visualisations can contribute to these problems. In this study we make a distinction between the technical experts and non-technical domain experts. The technical experts have knowledge of ML models, but no knowledge of the domain. In contrast, the non-technical experts who are specialised in the field have knowledge of the domain, but not of ML models. We have excluded technical domain experts, who have knowledge of both ML models and the domain. To discover the needs of these two groups, semi-structured interviews are conducted. More information about the protocol and sampling method can be found in Section 4.

- *Design*:
  During this phase, the visualization researcher initiates the design of the visualizations. The design requirements are derived from interviews and literature research on existing interfaces. Chapter 6 elaborates on the complete process, from designing low-fidelity prototypes to generating the final design.

- *Implement*:
  In this stage, the final design in the previous stage is implemented by the researcher.

- *Deploy*:
  In the final stage of the core phase, a tool is implemented and feedback is gathered through field testing. This research explores the development and evaluation of visualisations with actual users. Chapter 7 details this process.

1.5.3 Analysis Phase

The last category of this framework consists of two stages:

- *Reflect*:
  Reflection is a vital part of any research. A critical reflection should be carried out on the methods and findings of the research. The primary task in the reflection process is to properly describe the relationship of the research to the larger research area, and how the previously proposed design guidelines can be improved. In this research, the reflection is discussed on the basis of a discussion (Chapter 9) and conclusion (Chapter 10).

- *Write*:
  The final phase involves reporting the research. That is this written document for this research.

2. Background

The topic of this research lies at the intersection of three research domains: law enforcement, interpretable Machine Learning (iML), and data visualisation. For each of these three domains, this chapter presents key terminology and aspects used to describe the scope in which this research operates.

In this chapter, we examine sub-question "*SQ1.1: What methods are there to make ML models interpretable, particularly for law enforcement?*". This chapter will address the precondition phase of the nine-stage framework. First, Section 2.1 will provide an overview of the Dutch police, as this thesis is being carried out in collaboration with the National Policelab AI (NPAI). This overview will serve as the foundation for the winnow phase. Further details on the chosen target group for this research will be provided in Chapter 4. The subsequent sections will cover the main concepts of ML (Section 2.2) and visualisation (Section 2.4). This knowledge is part of the learn phase and will form the basis for the later stages of this research.

2.1 Dutch Law Enforcement

Law enforcement agencies are responsible for maintaining the rule of law, ensuring compliance with the law and providing assistance to people in need [1, 2, 3]. Every country has a unique law enforcement structure [4]. Therefore, this section will concentrate on the unique structure of law enforcement in the Netherlands, as this thesis is being developed in cooperation with the Dutch National police.

In the Netherlands, law enforcement agencies fall under the responsibility of the Ministry of Security and Justice [8, 5]. These agencies include many different services and institutions, including, for example, the Public Prosecution Service (in Dutch "Openbaar Ministerie"), the national police, and The Royal Netherlands Marechaussee (in Dutch "Koninklijke Marechaussee") [8]. Each of these services and institutions has its own role in society. For example, the Public Prosecution Service is the only body that can bring a suspect before a criminal court. It is responsible for the investigation and prosecution of criminal offenses [6].

The National police, on the other hand, has a special role in society. They have the constitutional right to use force to enforce the law [5, 6]. In addition to the National police, the Netherlands has a police force with military status, the Royal Netherlands Marechaussee [3]. They are responsible for the safety of Dutch airports, but can sometimes also be deployed in other places if necessary [3]. This research will focus on the Dutch National police.

The Netherlands has one national police force, consisting of ten regional units headed by a chief. These units are divided into districts and further into basic units. Furthermore, the national police force comprises specialist criminal investi-

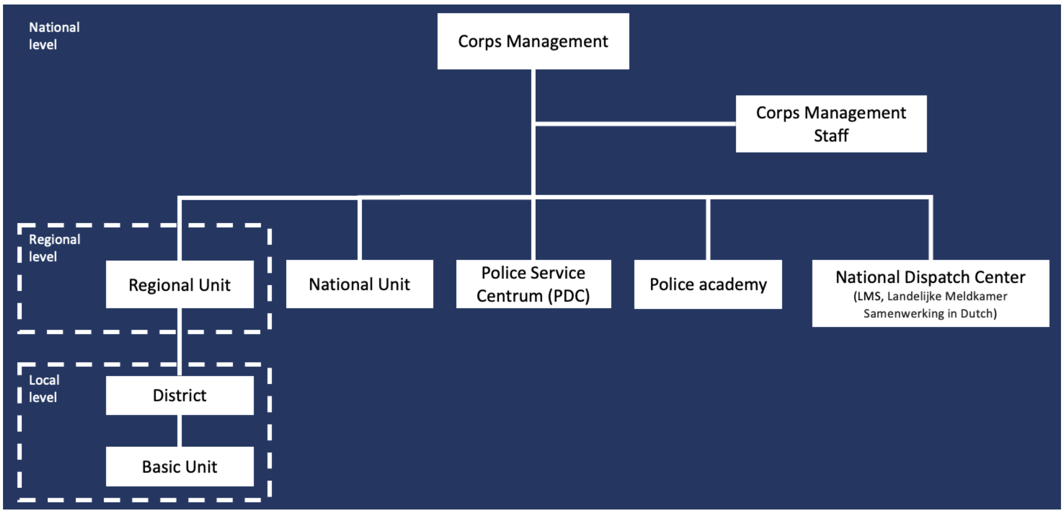


Figure 2.1: *Chart police organisation of Dutch National Police [38]*

gation units and royal and diplomatic security [3, 38]. The organisational chart is shown in Figure 2.1. The primary objective of the Dutch police is to guarantee that citizens abide by the law by acting against crimes and offences and by providing assistance and support to other emergency services (e.g. ambulance and fire brigade) [1].

2.1.1 Big Data in the Law Enforcement Domain

The police should possess the capability to analyse big data sets [12, 13] in order to discover hidden patterns and insights [15] in order to enforce the law. However, these datasets are of very large size and grow exponentially over time [12, 13]. Moreover, they encompass diverse modalities [9, 10, 11, 5], which makes the analysis of these datasets complex. For instance, it is possible for police data to contain text, audio, video, or images [9, 10, 11, 5]. Data can also be a combination of two or more modalities, for example, the information that investigators gather during the analysis of a suspect's laptop is worth examining. The laptop may contain emails (text), voice messages (audio) and images, and in order to get all the relevant information, the police need to analyse all these modalities found in the laptop.

In this research, we focused on the text modality, which the police predominately use to acquire and process information. For the police to function effectively, it is paramount that society perceives them as credible and has confidence in their conscientious execution of their responsibilities [38, 5]. To accomplish this, the police must ensure transparency to the public concerning their decisions. They achieve this by disclosing all their decisions and the findings of their inquiries [38, 5].

In summary, law enforcement agencies are dealing with big data that is extremely difficult to analyse due to the large, complex and unstructured nature of the data sets. In addition, these datasets are growing at an exponential rate, making them too complicated for human analysis and hindering the ability to make data-driven decisions. Consequently, automating the analysis of these datasets is crucial in enabling data-driven decisions to be executed. ML models are frequently utilised for such tasks, given their ability to guide important and complex human decisions [16] by identifying trends and patterns in large amounts of data [17].

2.2 Machine Learning

Machine learning is a field of research in which computers learn from models to generate behaviour or enhance predictions based on data. To accomplish this, an ML algorithm receives input from the data and generates an ML model. This model then transforms the input into a prediction, which is the ML model's estimation of the target value based on the given characteristics [32].

ML algorithms are a set of rules that a machine follows to achieve a specific goal. These algorithms are fed with training data and use this data to develop rules on how to make predictions or perform a particular task. These rules are learned so that new inputs can be converted into outputs. The result of the learning process is the ML model, which can then generate the desired output or predictions on new data [32], see Figure 2.2. One way to think of an algorithm is as a recipe that lists the input, the output, and the steps required to convert the input to the output. The model would then convert new inputs into an output based on the learned recipe.

An ML model can serve diverse functions, such as classification, regression, clustering, outlier detection, or survival analysis [32]. The function of a model depends on the dataset and the context in which it is used [32].

In law enforcement, ML models have many potentially useful applications, such as user-friendly services for citizens, personnel planning and crime prevention through predictive policing, improving productivity and paperless workflows, and processing large amounts of data [5].

However, the use of ML models in high-stakes decisions can have serious consequences for people's lives [5, 21, 40, 23]. For instance, consider the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), which is a model utilised in the US criminal justice system to predict a defendant's likelihood of recidivism [41]. This model was used to guide decisions about probation, parole and bail. Nonetheless, COMPAS does not explain its behaviour. However,

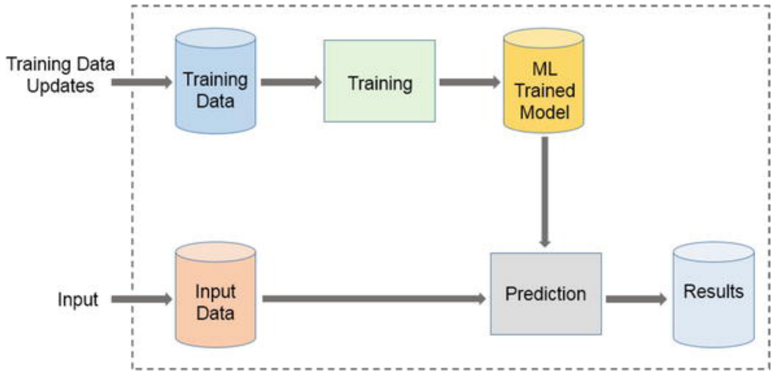


Figure 2.2: A machine learning workflow [39].

COMPAS has been accused of discriminating based on race and socioeconomic information (such as how often a person is paid below minimum wage) [40]. This discrimination has resulted in criminals with extensive criminal records being given a low COMPAS score and incorrectly labelled as low-risk, and vice versa [40]. Tragically, as a result of this bias, a very dangerous offender was released on bail. While on bail, he committed a murder [40].

One way of mitigating this behaviour is to increase the transparency of ML models and enable them to explain their behaviour to users and those affected by the models [42]. However, the opacity of some models can hinder the verifiability and transparency, and thus the accountability, of a decision. As a result, people may become overconfident, i.e. they accept biased decisions from an ML model without checking the behaviour of the model [43]. In other words, they take the output of a model as truth without knowing that it is actually true.

Explainable Artificial Intelligence (XAI) aims to tackle these problems, by creating techniques that produce more explainable models that are understandable to human users, while maintaining a high level of performance [44]. This term was coined by the DARPA program [44]. Since XAI techniques are bound to specific types of ML models we must identify the different types of ML available. This will enable us to select the appropriate XAI techniques for our problem domain.

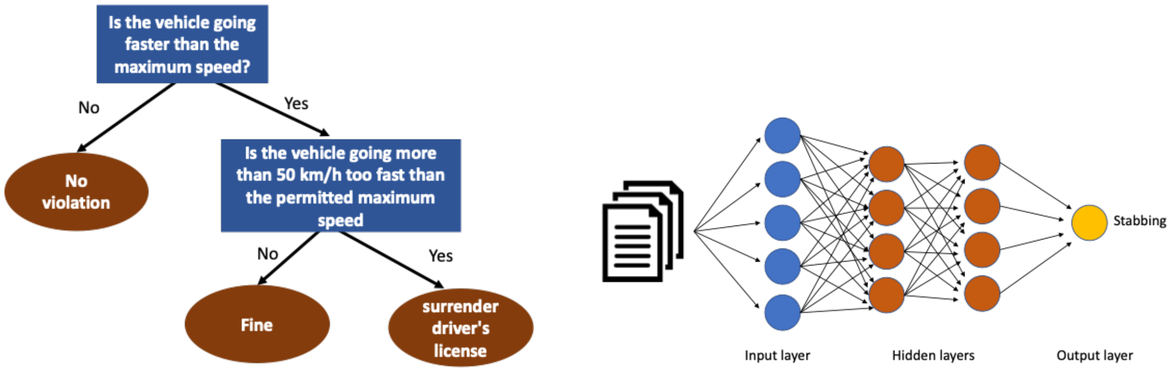
2.2.1 Types of ML Models

In this section, we will examine the different types of Machine Learning (ML) models available, with the aim of better understanding these models so that we can select the appropriate Explanable Artificial Intelligence XAI technique that best suits our problem domain. In the literature a clear distinction is made between transparent and black box models [45]. We will compare and contrast these two categories. First, transparent models and their advantages are presented. Next, we will present black-box models and their advantages.

Transparent models are considered transparent if it by itself is easy understandable for the human interacting with it [46, 32, 45]. The user is able to understand how the model behaves. In other words, the user is able to understand how the model transforms the input into an output. These types of models are also referred to as intrinsic models [32, 45].

Only a limited number of models are acknowledged for their transparency: rules, linear models and decision trees [47, 48]. However, only sparse versions (one with not too many features and/or decisions) of these are considered to be transparent [47, 48]. A decision tree with 1.000 nodes is still opaque, as it would be too large for a human being to examine.

Consider the small decision tree in Figure 2.3a as an example of a transparent model. This tree is a set of rules for determining whether a vehicle violates a speed limit. Suppose the input is a car going 30km/h on a road where the speed limit is 30km/h, this model would classify it as no violation. As a user, one can check if the model used the rules correctly.


Is the vehicle going faster than the maximum speed?
No
Yes
No violation
Is the vehicle going more than 50 km/h too fast than the permitted maximum speed
No
Yes
Fine
surrender driver's license
Stabbing
Input layer
Hidden layers
Output layer

(a) *A simple decision tree*

(b) *Deep Neural Network (DNN)*

Figure 2.3: *Examples of a transparent model (a) and a black-box model (b)*

Nevertheless, the small decision tree in Figure 2.3a has a low performance and accuracy on high dimensional datasets. The tree does not account for unusual circumstances, such as an ambulance, fire engine, or police car exceeding the speed limit while responding to an emergency call. One could add more rules to ensure the model applies to more situations. Consequently, this increases the complexity and reduces the user's understanding of the model. Eventually, the addition of an excessive number of rules makes human verification impossible, leading to an opaque model.

To improve performance and accuracy on high-dimensional datasets, highly complex and opaque models, called black-box models, are used. A model is considered a black-box model if it is (i) too complex for a human user to understand [45, 5, 32], or (ii) proprietary [23]. A black-box model is too complex if the user can observe the inputs and outputs, but does not have a meaningful understanding of the inner workings of the model. In other words, the user does not know how the model transformed the input into the output [45, 5, 32]. A model is proprietary if the user does not have access to the input [23]. Companies and organisations can have several reasons for not disclosing the input. For example, the police cannot share the inner workings of a model, because they store people's personal information such as home addresses, and citizen's service number. Note that one does not exclude the other: a model can be both too complex and proprietary [49].

An example of a black-box model is a deep neural network (DNN). The example in figure 2.3b shows a DNN classifying the topic of a report. In this model, your input is a report and the model outputs a topic. This topic is the most likely topic according to the model. In the example in Figure 2.3b, the model classified the topic of the report as stabbings. Both inputs and outputs are visible to human users, however, the model's reasoning is not. The reason for this may be that the human user does not have access to this information, because the model is proprietary, or because the human user does not know how the model reasoned. Consequently, comprehension becomes challenging for humans.

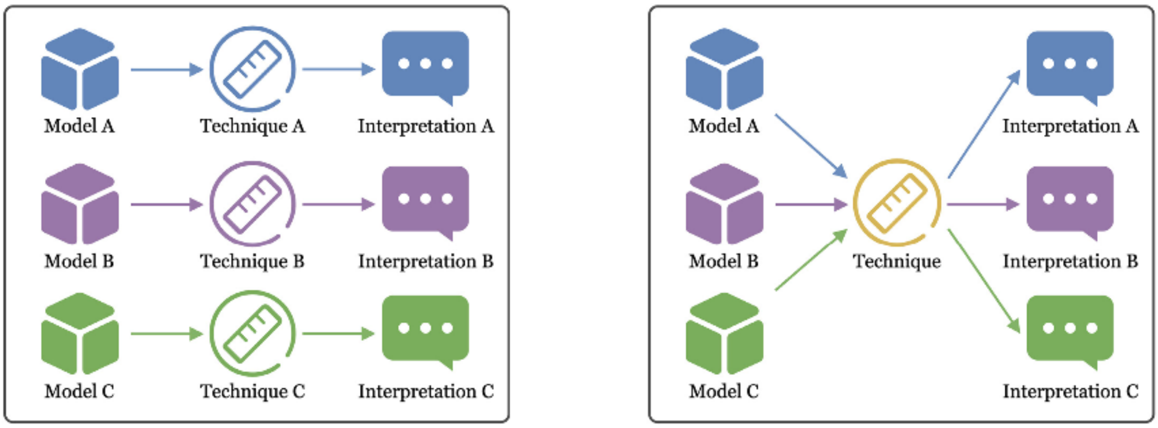
Suppose the decision tree in Figure 2.3a contains the brown round circles that symbolise the different outputs, but not the blue rectangles that symbolise the rules. Then we could give the model an input and it could generate an output for us. Nonetheless, as end-users, we would remain unaware of the rules applied by the model, and the model would be considered a black-box model.

Transparent models can be comprehended and interpreted by users. However, due to their inability to handle high-dimensional datasets with diverse cases, it is clear that such models are unsuitable for the law enforcement domain. On the other hand, black box models are too complex for human users to understand. In order for users to make informed decisions using these models, it is essential that the models are able to explain their behaviour.

2.2.2 Explaining Black-Box Models

Now that it is known that complex black-box models are the most appropriate for the law enforcement domain, as compared to transparent models, there is a need to find methods to explain them effectively in order to guide decision making. Fortunately, there are additional tools available for the explanation of complex black-box models. These techniques are also known as post-hoc [32, 45]. They are implemented after model training and can be divided into two distinct categories [32]:

- Model-specific methods: are methods that can be employed to explain particular types of ML models (Figure 2.4a). Such methods consider the precise characteristics and architecture of the ML model, thereby being constrained to certain types of ML models. Suppose this method is applied to explain neural networks, then it cannot be applied to regression.
- Model agnostic methods: are methods that are applicable to any type of ML model (Figure 2.4b). This method uses only inputs and outputs to explain an ML without understanding its internal structure. The advantage is that these methods are not restricted to specific types of ML methods. Therefore, in this research, we focus on this type of methods. The aim of



(a) *Model-specific* **(b)** *Model-agnostic*

Figure 2.4: *The difference between model-specific and model-agnostic interpretable MLtechniques [50].*

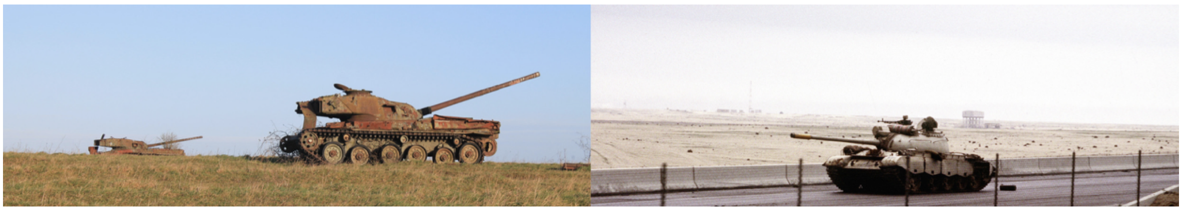


Figure 2.5: Examples of possible military tanks misclassified depending on the background, adapted from [49].

> our research is to create a design language that explains ML models to decision makers in the law enforcement domain, without being limited to the type of ML model. The research utilises a model agnostic approach resulting in the findings being applicable to explaining local ML models for decision makers in law enforcement beyond our target group.

An explanation is crucial not only for verifying that a model generates the accurate output but also for ascertaining if the model's behaviour is correct. In reality, a model may produce the correct output but exhibit undesirable behaviour by using incorrect or biased reasoning to transform the input into the output [40, 47, 51]. In his paper [47], Freitas illustrates such behaviour. The author outlines the application of a black-box classifier in a military scenario. The purpose of the model was to distinguish between a friendly tank and an enemy tank. Although the model performed with exceptional accuracy on the test set, the accuracy dropped dramatically when the model was used on new data. The model showed a bias; it classified based on the background colour, labelling a tank with a sunny background as friendly, while a tank with a cloudy background was considered hostile (Figure 2.5).

Ribeiro et al, [51] describe a similar example. The study shows that the presence of snow in the background (Figure 2.6) affects the ability of a model to classify an image as a husky or a wolf. Without the explanation presented in Figure 2.6, it is unfeasible to verify whether a model incorporates undesirable biases. Nevertheless, conducting such a check is crucial when employing models in high-stakes decision making scenarios, given the severe consequences of errors and biases [5, 21, 40]. Hence, it is crucial to provide an understandable and transparent explanation to decision makers to ensure the accuracy of a model's output and

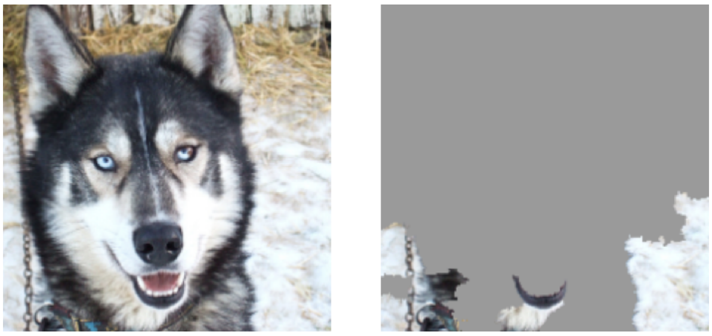


Figure 2.6: Examples of explanation of a model classifying an image containing either a wolf or husky, adapted from [51].

the soundness of the reasoning employed in obtaining it.

The example of the COMPAS model, in section 2.2, also showed that the explanation is crucial to check whether the model does not contain any unwanted bias. Now COMPAS is accused of discriminating based on race and socioeconomic information [40].

However, there are two types of explanations, global and local [32, 19]. Global explanations clarify how a general ML model functions, while the focus of this dissertation is on local explanations. Local explanations explain specific input-output pairs, that is, they detail how the ML model operates for a single prediction.

2.3 Designing Explanations

So far we have seen that black-box models are commonly applied as a type of model within law enforcement. We have also seen that it is important to explain these black-box models so that the model does not contain unwanted biases and is accurate. By using post-hoc methods we can explain the black-box models to the user. However, we must ensure that the explanation is designed to meet the needs and requirements of the explainee.

This section is divided into several subsections. In subsection 2.3.1 we present several roles that are able to interact with an ML model. Following this, Subsection 2.3.2 describes how not only the role but also the characteristics of the user influence the desired visualisation. Subsection 2.3.3 discusses how we can design explanations. By exploring these topics, the reader will gain insight into how we can design an explanation of a ML model for the target user so that the explanation meets the user's needs and requirements.

2.3.1 Explaining for Different Roles

XAI systems describe their reasoning behind a decision, by explaining the behaviour of an ML model to the end users in an explanatory way [19]. However, different users require different types of explanations [22]. Rather than asking if the model is explainable, interpretable, or trustworthy, we might better ask to whom the model should be explainable, interpretable, or trustworthy [31, 52].

According to Tomsett et al. [31], agents (whether humans or machines) may undertake six distinct roles in their interaction with an ML model (as seen in Figure 2.7). These roles are not mutually exclusive, meaning that an agent may perform multiple roles. It is vital to understand the way in which these roles interact with an ML model, as we are specifically interested in explaining ML models to decision makers. The different roles are:

- Creators: these agents create the ML model.
- Examiners: these agents investigate and audit the ML model.
- Operators: these agents provide the model with inputs and receive the outputs provided by the model. The operators interact directly with the model.

- Executors: these agents receive information about the outputs of the model from the operators and make a decision based on this output. The executors interact indirectly with the model.
- Decision-subjects: these agents are affected by the decision that the executors make.
- Data-subjects: the data of these agents are used to train the ML model.

Let us now sketch a scenario within the police force that shows which roles can possibly be filled by which persons. Note that in alternative scenarios the roles can be filled by different people, but they can also be filled by the same people.

Scenario: The police develop and use a model to determine if someone, based on earlier occurrences of shoplifting, has committed shoplifting. In this scenario, the roles can be divided as follows:

- **Creators:** the data scientists, data collectors, product managers and developers are responsible for creating the model. The data collectors provide the data, the data scientists develop the model with this data, and the product managers and developers ensure that the model is implemented in such a way that the executor can start working with the model.
- **Examiners:** the chief officer in charge of the department or district. This department or district is the department or district in which the model is used. For example within a district such as City Utrecht.
- **Operators:** the data scientist.
- **Executors:** the officer who must determine whether the person will be ticketed for shoplifting yes or no.
- **Decision-subjects:** person suspected of shoplifting
- **Data-subjects:** persons who have ever been convicted of shoplifting and whose details have been recorded.

When designing an interpretable model, it is important to know the motivational goal for interpretability. Especially in real-world applications, it is important to

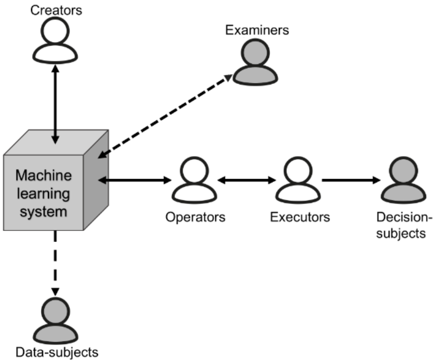


Figure 2.7: *Illustration of how different types of agents interact with an MLmodel, according to Tomsett et al., [31].*

consider the necessity of this [45], as biases can have major consequences [5, 23]. Different roles have different goals and should therefore receive different explanations, tailored to that goal. For instance, the executor is responsible for making decisions related to individuals and assessing the validity and absence of biases in the model's output. Conversely, the creator has to evaluate whether the model performs adequately or whether something needs to be modified, and is therefore concerned with a completely different set of questions.

2.3.2 Different Characteristic Needs

Not only the roles that agents have in relation to a system influence the necessary type of explanation, but their personal characteristics also play a part [25, 53]. For instance, an agent may only possess domain knowledge but lack the required technical knowledge. In such cases, one must ensure that the explanation is not overly technical, while still being understandable to the agent [31]. Agents in diverse roles possess distinct backgrounds, experiences and technical or domain expertise [25]. Consequently, both the role and personal characteristics have an impact on the method of conveying explanations.

In this research, we will refer to four different types of users, limiting ourselves to human users:

1. **Domain experts:** refers to agents who are experts in the problem domain but do not have sufficient expertise on ML. In the context of the police, it entails understanding all terminology used within the field.

2. **Technical experts:** refers to agents who are experts on ML but not in the expert domain.

3. **Technical domain experts:** refers to agents who are experts on ML and in the problem domain.

4. **Novices:** refers to agents lacking expertise in both the problem domain and in the field of ML.

There are several reasons why an explanation is necessary. However, how do we know what to explain? According to Mohseni et al., there are six common types of explanations used in XAI systems [19]:

- **How explanations:** explain how the model works [19].

- **Why explanations:** explain why a decision was made for a particular input [19].

- **Why-Not explanations:** explain why a certain decision was not in the output of a system [54].

- **What-if explanations:** demonstrate how the model output is affected by data changes and different algorithms by new inputs [55], manipulations of the inputs [56], or changes in the model parameters [57].

- **How-to explanations:** describe how hypothetical adjustments of the input can result in different outputs [56, 58].

- What-else explanations: present similar instances of input to the user that generate the same or similar outputs [59].

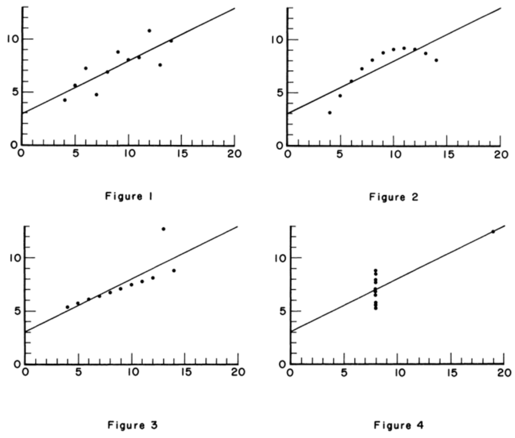
In summary, the type of explanation depends on the role the person has towards the model and what question they are trying to answer. But how can we make these different explanations? In the next section, we will look at how to create explanations.

2.3.3 How to Explain?

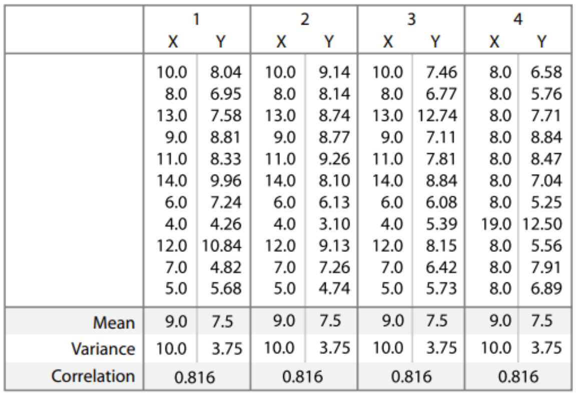
Up to now we have seen that there are different needs for an explanation of a model for different roles. Not only the role, but also the characteristics of a user, such as knowledge about a domain or about ML models, play a role. How can we effectively generate an explanation that meets the needs of our target audience? In this subsection, we will explore methods for providing model explanations.

Explanations can be designed using a variety of different modalities, e.g., verbal, visual, and numerical elements [19]. Verbal explanations use natural language as phrases and words to describe the workings of a model, whereas visual explanations use visual elements to describe the workings, and numeric explanations make use of numerical metrics [19]. In addition, a combination of these three different modalities can also be used [19]. This research mainly focuses on visual explanations, as these explanations can depict the data in more detail.

Sometimes it is necessary to display more details, as summaries can cause information loss. Anscombe's Quartet is an example, designed by the statistician Anscombe, which depicts that a summary can be an oversimplification of the dataset [60]. Anscombe's quartet is constructed to illustrate the importance of plotting data before analysing it. The quartet consists of four datasets that have the same statistical observation regarding the variance, mean, and correlation (Figure 2.8b). The table, in Figure 2.8b, with the four datasets would therefore suggest that the datasets are identical. Upon visual inspection, it is evident that



(a) *Four different XY plots*



	1		2		3		4	
	X	Y	X	Y	X	Y	X	Y
	10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
	8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
	13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
	9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
	11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
	14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
	6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
	4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
	12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
	7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
	5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89
Mean	9.0	7.5	9.0	7.5	9.0	7.5	9.0	7.5
Variance	10.0	3.75	10.0	3.75	10.0	3.75	10.0	3.75
Correlation	0.816		0.816		0.816		0.816	

(b) *Four datasets with identical descriptive statistics*

Figure 2.8: *Anscombe's quartet of four different XY plots of four datasets which have identical descriptive statistics (identical means, variance, correlation, and linear regression lines), adapted from [60, p 19-20]*

the data sets are not identical because:

- The first scatter plot (above left) shows a simple linear relationship with scatter; both variables could be normally distributed.
- In contrast, the second scatter plot (above right) certainly does not show a normal distribution. Although a clear correlation exists between the variables, it is not linear.
- The third scatter plot (bottom left) shows a strong linear correlation with a large outlier.
- Similarly, the fourth scatter plot (bottom right) shows an outlier.

2.4 Visualisation

In this section, we discuss how to design visualisations. The section is divided into two subsections, each focusing on a specific aspect. In Section 2.4.1, we describe what the definition of the term visualisation is. Section 2.4.2 discusses the methods to design effective visualisations.

2.4.1 Defining Visualisation

Before exploring the potential choices for visual explanations, it is good to understand what is meant by the term visualisation. Stuart et al. [61, p. 6], define the term visualisation as: *"The use of computer-supported, interactive, visual representations of abstract data to amplify cognition."* Whereas Munzer [62, p. 1] defines it as *"Computer-based visualisation systems provide visual representations of datasets designed to help people carry out tasks more effectively."* We define the term visualisation, by employing the key insights of the above-stated definitions, in this thesis as follows:

Visualisation refers to a computer-based interactive visual representation of abstract data, that amplifies cognition and helps humans carry out tasks more effectively.

This definition implies the following:

Abstract data: refers to data with no obvious or natural visual representation. This also includes, for example, unstructured data, which the police use.

Interactive: refers to how a user can change what and how something is visualised. This is an important aspect since each static visualisation can only answer sub-questions of a problem. Changing what and how something is visualised can help to answer more questions [63].

Human user: if there is no need for a human judgment, then there would also be no need for a visualisation.

Amplify cognition: refers to the ability to solve a problem with less effort, in a shorter time and with more accuracy. This is important for the police, because the police have to make quick and accurate decisions.

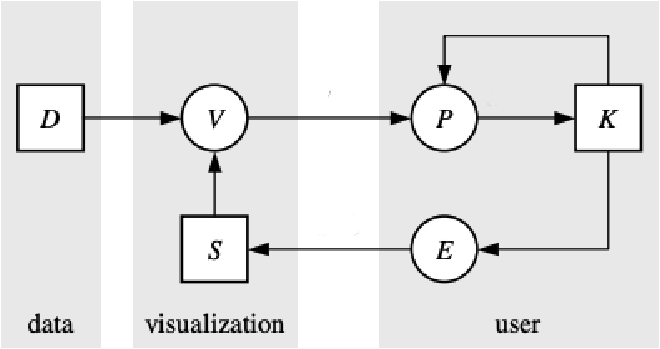


Figure 2.9: Illustration of the context in which a visualisation operates, adapted from [64, p 2]. Data D is being transformed into a visualisation V according to a specification S. This visualisation is dependent on the interpretation of the user, which is influenced by the user cognitive and perceptual properties P, the users knowledge K and the types of interactions E the user use to adapt the appearance of the visualisation.

Visual representation: increases people's knowledge about a certain phenomenon and maximizes the changes that this knowledge is truthful, as we saw in the example by Anscombe [60] (Figure 2.8).

How a visualisation is interpreted depends very much on the context in which it is placed [64]. For example, the context in which a visualisation is used can differ between scientific and informational contexts, thereby dictating the optimal approach for creating such visualisations. Consider the example: an information graphic can be created for the general public, guaranteeing comprehension by all.

The context in which a visualisation works is illustrated in Figure 2.9. Visualisation (V) is observed by a user after which it gains knowledge (K) of this visualisation. The amount of knowledge gained (K) depends on the image (V), the user's current knowledge, and the specific properties of the user's perception and cognition (P) [64].

As far as the influence of K is concerned, a data scientist will be able to extract more information from a technical visualisation than a layman. But even if a lot of knowledge is available, the extra knowledge that the image shows can be low. This can occur, for example, when you show the map of the Netherlands with the provinces to a Dutch person compared to a foreigner. The foreigner will then learn more from the map than the Dutch person [64].

A user's perception and cognition also greatly influences how a visualisation is perceived and whether knowledge can be extracted from the visualisation. A colorblind person will be less effective at extracting knowledge from a colorful visualisation than someone with full vision. In addition, some people are better at recognising patterns and structures and this does not always immediately pop out of an visualisation. To ensure that people with different cognitive aspects can perceive an visualisation, interactions (E) are often provided. This allows a user to adjust the way the visualisation is displayed in order to extract the knowledge from the visualisation [64].

To ensure that the visualisation is designed for the user, we need to know in which ways we can design a visualisation. We will discuss this in the next section.

2.4.2 Creating Visualisations

To create visualisations, visualisation techniques are used. They use a combination of two aspects: (1) graphic elements called marks, and (2) channels that can change the appearance of the marks [62].

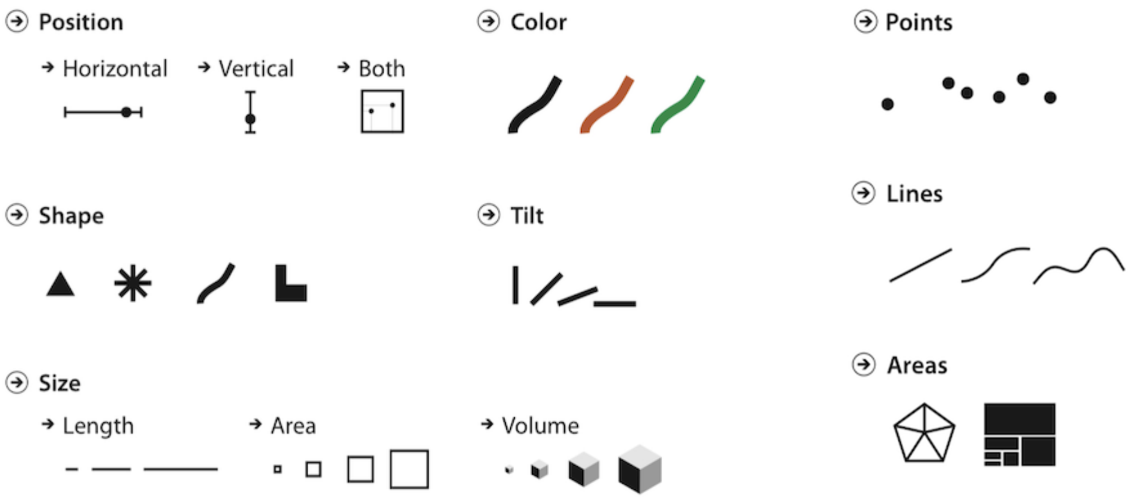
Marks can be classified according to their number of spatial dimensions. The most commonly used ones are: zero-dimensional (points), one-dimensional (lines), and two-dimensional (areas), see Figure 2.10b [62].

Channels can have many different forms, see Figure 2.10a. They can be categorised into two types [62]:

1. *Identity channels:* describe what or where something is, e.g., position.
2. *Magnitude Channels:* describe how much there is of something, e.g., size.

Which visualisation techniques are used depends on the type of represented data, such as numbers, text, multidimensional, or networks [63]. This research will focus on text visualisation techniques since this is the type of data that is mostly used by police officers. Text visualisation techniques create visualisations for (raw) text data, or results of text mining algorithms [65].

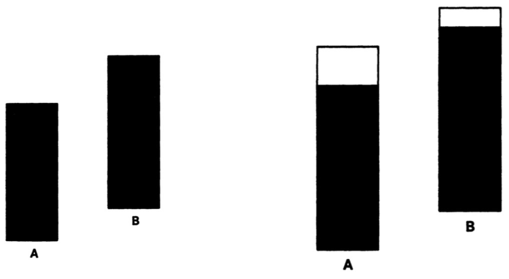
Visualisation design should be guided by the principles of expressiveness and effectiveness [62]. There are many different marks and channels and not every combination is possible. According to the expressivity principle, the visual coding should express all and only the information in the dataset attributes [62]. For example, ordered data must be represented in a way that our perceptual system intrinsically perceives it to be ordered, and disordered data must be represented



(a) *Channels*

(b) *Marks*

Figure 2.10: *The different types of channels and marks [62].*



(a) *Unframed and unaligned bars*

(b) *Framed and unaligned bars*

Figure 2.11: *Example of Weber's Law [62, 66]*

so that we perceive it as disordered.

According to the effectiveness principle, the most important attributes should stand out the most, and the least important the least [62]. There are several aspects that can influence the effectiveness of a visualisation (accuracy, discriminability, separability, popout, and grouping). Besides these two aspects, we also need to consider Weber's Law, which states that we judge the difference between two items based on relative and not on absolute difference. For example, it is much harder to compare the two bars in Figure 2.11a than comparing the two bars in Figure 2.11b [66].

Therefore, determining whether the channels and marks are used effectively is important when designing visualisations. One cannot use the same channels for different parts of a visualisation. To account for this, the following ranking has been created, with the most effective channel per data type (ordered vs. categorical) listed at the top:

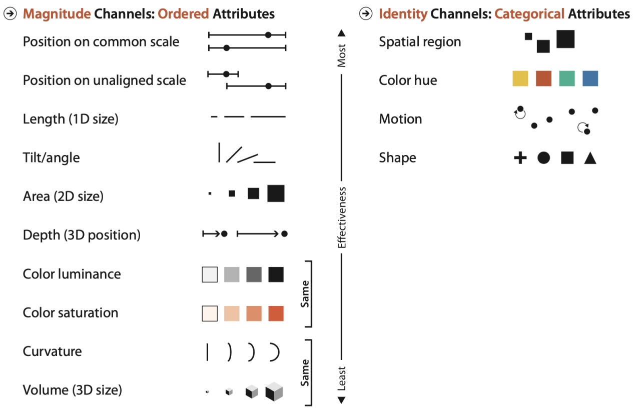


Figure 2.12: *Ranking of channels according to their data type [62].*

Datasets are sometimes large and complex that displaying everything at once would lead to information overload [62]. To reduce this, one can use different actions to navigate through a visualisation. These actions are referred to as interactions. Interactions are defined as the interplay between a human and a data interface with a data-related intent, where the human performs at least one action to which the interface responds and that response is observed by the human [67].

The two most used actions in a 2D visualisation are zooming and panning [62]:

- Zooming is the action that allows you to view more or less detail [62]. By zooming in, one goes deeper into the records, seeing less or a single element more closely. Zooming out allows one to see more elements, but in less detail.

- Panning is the action in which you move the image along the horizontal axis of the image [62]. For example, you pan to the left or right if you are zoomed in and want to see items that occur earlier or later.

One can also select an item in the view for further inspection. A selected item must provide feedback to the user so that it matches the user's intent. What is often done is to highlight the selection. For example, when someone presses a button, that button gets a different colour.

The use of multiple views is another way to reduce complexity [62]. The strength lies in the fact that it makes it easier for the user to compare different details in the data set. They do not have to remember what they saw in a different part of the dataset before.

Summary

In this chapter we have seen that the use of ML models can play an important role in decision making process in the law enforcement domain. However, we have also seen that explanations are very important to avoid errors and unwanted bias. To ensure that the user can make a decision based on the output of the model, it is important that ML models can explain their behaviour to the user. There are different roles that interact with a model, so it is important that we can validate that the explanation of the behaviour of the model is designed for the right role and intent. We have also seen that visual explanations are an effective way to design explanations. However, there are several principles to consider so that the user can interpret the visualisation as intended. To reduce information overload, one should also add different ways to interact with the visualisation, give the option for multiple views and select items of interest.

3. Literature Review

There are various techniques and approaches to visualising Machine Learning (ML) models for explanatory purposes, as noted in chapter 2. Tailoring the design to the audience is critical, as users often seek different levels of information when interacting with models. Since we are interested in exploring how to visualise local explanations for decision makers. The objective of this section is to recognise existing tools that explain ML models to domain experts, and to see how the visualisation properties of these tools can be applied to our research. The findings of this review should provide an answer to sub-question "*SQ1.2: How are these methods visualized in existing tools relevant to law enforcement?*", which form a foundational element of our design. The interview data, which will be discussed in the next chapter, will complete this foundation.

So far, we have found that black-box models are the most appropriate type of ML models to use in the law enforcement domain. This is because law enforcement requires the analysis of high-dimensional data sets. Transparent models have been found to present deficiencies in this case. Nevertheless, it is imperative that we ensure that the black-box models can explain their behaviour using post-hoc methods.

Various post-hoc methods are available [32]. However, this study will concentrate on local explanations, which already eliminates some of these methods. Moreover, there is no particular ML model on which we focus at this point. Therefore, we require model-agnostic post-hoc methods that can explain ML models locally.

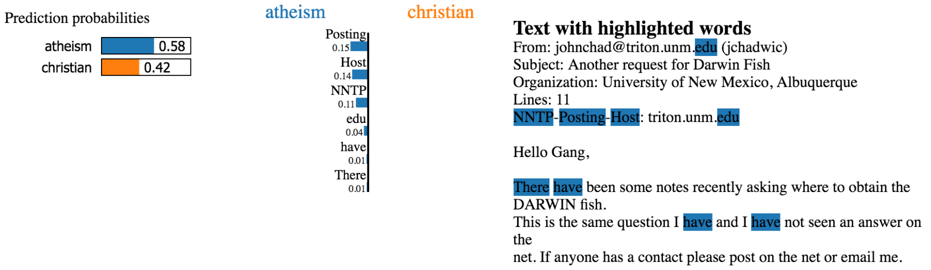
The two most well-known model-agnostic post-hoc methods in the XAI field are [21] Local Interpretable Model-Agnostic Explanations (LIME) [51] and SHapley Additive exPlanations (SHAP) [68]. Our research will be centred on these two methods exclusively for clarity purposes. Firstly, we will elaborate on these two methods in Section 3.1, along with visualizations used by their developers. Then, Section 3.2 will provide a literature review of the domains in which these two methods are used to explain ML models and how they are visualised in the tools.

3.1 LIME and SHAP

LIME [51] and SHAP [68] are model agnostic and can explain decisions of ML models locally. In this study, we consider LIME and SHAP as relevant methods because the model agnostic nature of these two methods allows XAI developers and designers to provide a unified interface for the law enforcement domain, even if the ML model or the underlying data changes [32]. This section will explain these two methods.

LIME locally generates linear models around a single prediction to explain the models decisions locally. LIME can do this for data in text format, but also for images and tabular form [51]. LIME results in a list, disclosing the contribution of each feature to a data sample's prediction. This enables you to identify the most influential features. In text format, a feature can be a word, part of a word, part of a phrase or an entire phrase. The data scientist determines the format of the features.

Figure 3.1. portrays an example of how LIME classifies text according to two classes. In this example, the topic of an email needs to be categorized as either atheistic or Christian. Blue highlights indicate a word classified as atheistic, while orange highlights indicate words classified as Christian.



Text with highlighted words
From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the
DARWIN fish.
This is the same question I have and I have not seen an answer on
the
net. If anyone has a contact please post on the net or email me.

Figure 3.1: *An example of a LIME explanation for text classification [51].*

The bar chart in the middle of Figure 3.1 shows the top 10 features that contribute the most to the classification. Note that only the top 10 features are shown in the default visualisation. It is worth noting that other features may have an impact, yet they are never visualised in the default visualisation. In this example, the Posting feature has the highest contribution to the classification of the email.

The developers of LIME implemented a standard visualization technique to visualise the outcomes of the prediction. The visualisation consists of three different elements: on the far left is the legend showing what each colour in the visualisation indicate, in the middle is a bar chart showing the list of the top 10 most contributing features, and on the far right is the text being classified.

The text on the right remains in its original state. However, the keywords that contribute to the classification of the text are now highlighted in the relevant colour of the class.

SHAP calculates the feature importance of an individual prediction using Shapley values [68]. Shapley value is a concept from game theory that shows the average contribution for each player. In SHAP, the idea of Shapley values is applied to the contributions of features rather than that of players. For example, SHAP can be used to classify pieces of text, in which each word is an individual feature that contributes to the classification of the entire sentence.

An example of how the visualisation of these Shapley values in SHAP look like can be found in Figure 3.2. In this example, a sentence about a movie is classified as either a positive or a negative review. The positive features are displayed in red, whereas the negative features are displayed in blue. In this case, the features are whole phrases. The size of each bar is related to its contribution

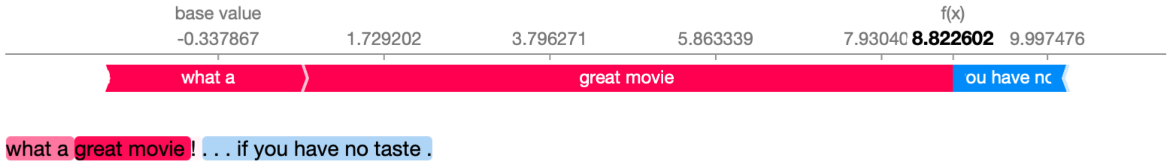


Figure 3.2: An example of a SHAP explanation for text classification [68].

to the classification of the piece of text. So in this example in Figure 3.2, great movie has the largest contribution of the two positive features that are visible.

Whilst LIME and SHAP have similarities, they differ in their techniques for explaining Machine Learning models. For instance, both methods utilise features, but LIME used word-level tokens and SHAP used whole phrases. In addition, there are also slight variations in the way on how they calculate each feature contribution. Given the technical nature of these calculations, they will not be further explored in this research. Furthermore, the visualisation methods used to provide the explanations of ML models also differ.

Although different, both LIME and SHAP provide explanations through an information-intensive visualization. These visualizations lack interactivity and guidance for users to better interpret the visualisations [69]. A study conducted by Kaur et al. [70], demonstrates that even technical experts may struggle to comprehend these visualizations to understand the explanation of the ML model. As a result, users may blindly rely on the output of a model even though this output may be biased [69, 70]. This situation can have major consequences if this happens in high social impact situations [5, 23].

To design a visualization that effectively visualises the local explanations of an ML model, we need to ensure that the user can correctly interpret the output and the local explanation of the ML model. The standard LIME and SHAP visualisations alone are not 100% capable of achieving this. Even ML experts can misinterpret these visualizations, despite their familiarity with ML models [70]. Therefore, we will investigate in the following section whether and how current interfaces employ these two methods to explain ML models to users with no prior experience in the field of ML.

3.2 Existing Tools

In this section, we present our literature review on the use of LIME [51] and SHAP [68] in tools for explaining ML models. Our objective is to identify the domains and target audiences for which these tools have been developed, and the visualisation techniques and methods used to visualise LIME and SHAP. Section 3.2.1 provides an overview of the methodology employed for literature selection. We will subsequently present our findings in Section 3.2.2. Finally, Section 3.2.2 will provide recommendations for our design based on this research.

3.2.1 Paper Selection Process

To identify relevant articles, we conducted a keyword search on Google Scholar. We focused our search on visualisation design and selected articles from three major visualisation platforms: TVCG, CGF, and CHI. It should be noted that this search was not a comprehensive systematic review, therefore it is possible that interfaces that can be found in the literature will not be dis- cussed.

We examined articles using the keywords "LIME" and "SHAP". Subsequently, we merged these two keywords with the keyword "law enforcement". These keywords could be found throughout the text. Articles that only described these methods, i.e. that did not use these methods in a tool, were removed. Our search employed the exact keyword method, ensuring that articles with terms like "shape" instead of "SHAP" were disregarded. We subsequently included the complete versions of the keywords, as we discovered that some articles on LIME only discussed the colour lime, adding noise to our results. Hence, we added the terms "Local Interpretable Model-Agnostic Explanations" and "SHapley Additive exPlanations". In addition, the key word "visualisation" was included for the CHI venue in the search, as it is not a venue that focuses on visualisation. The following section will analyse and examine the chosen articles in this literature review.

In the upcoming section, we will analyse and discuss the papers selected for this literature review.

3.2.2 Results Literature Review

In this section we will discuss which existing tools exist that use model-agnostic methods LIME [51] and SHAP [68] to explain the behaviour of an ML model to the user.

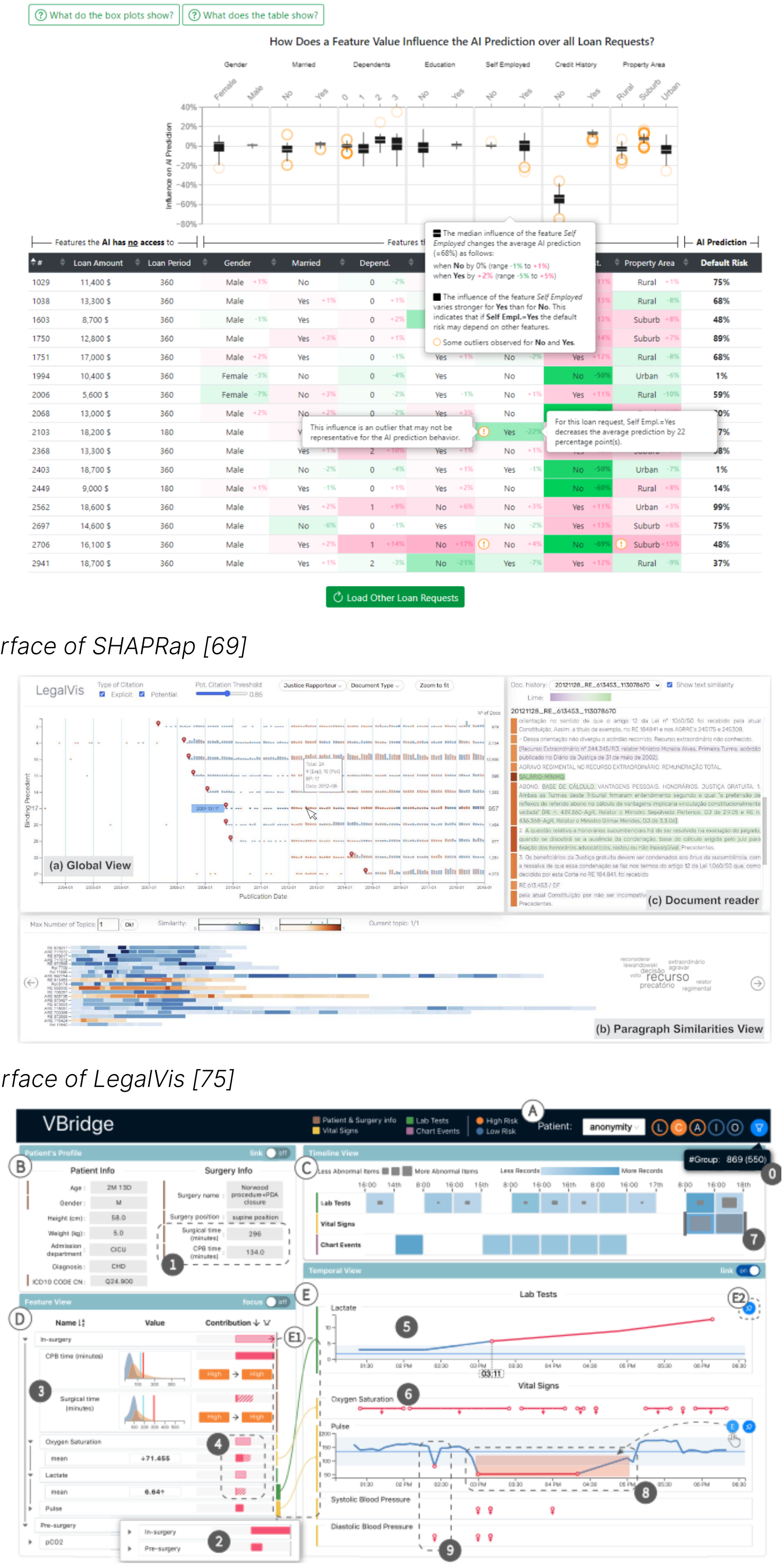
Target Group Explainable Tools

Many existing interfaces have been developed to explain the behaviour of an ML model to technical domain experts [71, 72, 73, 74]. However, these are often focused on how to develop and improve a model rather than on making a decision with the output of a model [26, 28].

There are a multitude of decision making tools, including ShaPRap [69], which focus on guiding human decision makers through intelligent and interpretable systems, see Figure 3.3. However, this particular study centers on decision makers who lack technical and domain-specific knowledge. As such, the tool was tested through a loan application decision, chosen by the authors due to their belief that anyone can make this type of decision.

Nevertheless, there are also a number of tools designed for non-technical domain experts, such as LegalVis [75], which is specifically developed to support legal professionals in analysing legal documents that reference, or could reference, binding precedents. The system assists domain experts in data analysis, although its purpose is to showcase pertinent documents rather than facilitate decision making.

Vbridge [76], on the other hand, is a tool designed for healthcare domain experts to support clinicians in making clinical decisions through forward and backward analysis workflows. The study demonstrated that linking model statements with



What do the box plots show?
What does the table show?
How Does a Feature Value Influence the AI Prediction over all Loan Requests?
Gender
Married
Dependents
Education
Self Employed
Credit History
Property Area
Female
Male
No
Yes
0
1
2
3
Rural
Suburb
Urban
Influence on AI Prediction
40%
20%
0%
−20%
−40%
−60%
−80%
Features the AI has no access to
Features th
AI Prediction

■ The median influence of the feature *Self Employed* changes the average AI prediction (=68%) as follows:
when **No** by 0% (range -1% to +1%)
when **Yes** by +2% (range -5% to +5%)

■ The influence of the feature *Self Employed* varies stronger for **Yes** than for **No**. This indicates that if **Self Empl.=Yes** the default risk may depend on other features.

○ Some outliers observed for **No** and **Yes**.

#	Loan Amount	Loan Period	Gender	Married	Depend.	[illegible]	[illegible]	t.	Property Area	Default Risk
1029	11,400 $	360	Male +1%	No	0 -2%	[illegible]	[illegible]	[illegible]	Rural +1%	75%
1038	13,300 $	360	Male	Yes +1%	0 +1%	[illegible]	[illegible]	[illegible]	Rural -8%	68%
1603	8,700 $	360	Male -1%	Yes	0 +2%	[illegible]	[illegible]	[illegible]	Suburb +8%	48%
1750	12,800 $	360	Male	Yes +3%	0 +1%	[illegible]	[illegible]	[illegible]	Suburb +7%	89%
1751	17,000 $	360	Male +2%	Yes	0 -1%	Yes +1%	No -2%	Yes +12%	Rural -8%	68%
1994	10,400 $	360	Female -3%	No	0 -4%	Yes	No	No -50%	Urban -6%	1%
2006	5,600 $	360	Female -7%	No +3%	0 -2%	Yes -1%	No +1%	Yes +11%	Rural -10%	59%
2068	13,000 $	360	Male +2%	No +2%	0 -2%	Yes +3%	No	[illegible]	[illegible]	[illegible]
2103	18,200 $	180	Male	[illegible]	[illegible]	[illegible]	ⓘ Yes -22%	[illegible]	[illegible]	[illegible]
2368	13,300 $	360	Male	Yes +1%	2 +10%	Yes +1%	No +1%	[illegible]	[illegible]	[illegible]
2403	18,700 $	360	Male	No -2%	0 -4%	Yes +1%	Yes -1%	No -50%	Urban -7%	1%
2449	9,000 $	180	Male +1%	Yes -1%	0 +1%	Yes +2%	No	No -60%	Rural +8%	14%
2562	18,600 $	360	Male	Yes +2%	1 +9%	No +6%	No +3%	Yes +11%	Urban +3%	99%
2697	14,600 $	360	Male	No -6%	0 -1%	Yes	No -2%	Yes +13%	Suburb +6%	75%
2706	16,100 $	360	Male	Yes +2%	1 +14%	No +17%	ⓘ No +4%	No -69%	ⓘ Suburb +15%	48%
2941	18,700 $	360	Male	Yes +1%	2 -3%	No -21%	Yes -7%	Yes +12%	Rural -9%	37%

This influence is an outlier that may not be representative for the AI prediction behavior.

For this loan request, Self Empl.=Yes decreases the average prediction by 22 percentage point(s).

(a) *The interface of SHAPRap [69]*

Doc. history: 20121128_RE_613453_113078670 ☑ Show text similarity

Lime:

20121128_RE_613453_113078670

orientação no sentido de que o artigo 12 da Lei n° 1060/50 foi recebido pela atual Constituição. Assim, a título de exemplo, no RE 184841 e nos AGRRE's 245175 e 245308.

- Dessa orientação não divergiu o acórdão recorrido. Recurso extraordinário não conhecido.

(Recurso Extraordinário n° 244.345/RJ, relator Ministro Moreira Alves, Primeira Turma, acórdão publicado no Diário da Justiça de 31 de maio de 2002].

AGRAVO REGIMENTAL NO RECURSO EXTRAORDINÁRIO. REMUNERAÇÃO TOTAL.

SALÁRIO-MÍNIMO.

ABONO. BASE DE CÁLCULO. VANTAGENS PESSOAIS. HONORÁRIOS. JUSTIÇA GRATUITA. 1. Ambas as Turmas deste Tribunal firmaram entendimento segundo o qual "a pretensão de reflexos do referido abono no cálculo de vantagens implicaria vinculação constitucionalmente vedada" [RE n. 439.360-AgR, Relator o Ministro Sepúlveda Pertence, DJ de 2.9.05 e RE n. 436.368-AgR, Relator o Ministro Gilmar Mendes, DJ de 3.3.06].

2. A questão relativa a honorários sucumbenciais há de ser resolvida na execução do julgado, quando se discutirá se a ausência da condenação, base de cálculo erigida pelo juiz para fixação dos honorários advocatícios, restou ou não inexeqüível. Precedentes.

3. Os beneficiários da Justiça gratuita devem ser condenados aos ônus da sucumbência, com a ressalva de que essa condenação se faz nos termos do artigo 12 da Lei 1.060/50 que, como decidido por esta Corte no RE 184.841, foi recebido

RE 613.453 / DF

pela atual Constituição por não ser incompatív

Precedentes.

(b) *The interface of LegalVis [75]*

(c) *The interface of VBridge [76]*

Figure 3.3: *The most promising tools for our research.*

patients' situational data via visual aids can enhance clinicians' interpretation of model predictions and their use in clinical decision making. Another tool designed for non-technical domain experts is Riseer [77], which identifies the evolutionary patterns of RIS and enables inter-regional comparisons based on visual designs to assist domain experts.

Although various tools have been created to aid domain specialists in decision making, none have been developed to support police officers in making law enforcement decisions. However, according to the researchers behind Legalvis, the tool can be used in other areas, including medicine and public purchasing [75]. This tool proves to be highly valuable in the field of law enforcement, since it can guide the process of finding relevant records. In the following subsection, we will examine crucial aspects of a visualisation that facilitate decision making.

Important Visual Aspects

A recurring theme in interfaces is the use of the visualisation mantra *"first overview, zoom and filter, then details-on-demand"* [66]. This is frequently employed in interfaces designed for the non-technical users. Examples include LegalVis [75], Vbridge [76], PromotionLens [78], and Oui [74]. Applying this principle has a significant benefit as it presents a comprehensive view of the entire data set [76]. This is important in the Vbridge tool because clinicians base their final decisions on raw data, which is necessary when making decisions that involve human lives because the wrong decision can have significant consequences [76]. It is also a crucial aspect of our research domain as decisions can have significant consequences on people's lives.

Moreover, the study conducted by Resk et al. [75], revealed that the consistent application of colour enhanced the interface's efficiency for tool users. These findings highlight the significance of the visualisation mantra, which facilitates the identification of specific inputs. Additionally, the ability to filter and search for pertinent documents was deemed highly practical. Our research also acknowledges the importance of this feature, as it aims to visualise local explanations of an ML models. Our users seek specific input-output pairs. However, the interpretability of LIME explanations by users was not evaluated in this study [75]. Hence, it remains uncertain whether highlighting features is an effective approach to explain the behaviour of ML models.

Most tools visualise the local explanations of an ML model [75, 76, 77, 78, 74]. However, ShaPRap [69] also visualises the global explanations. Yet it turns out that users often need more explanation than just the visualisation, which shows features that contribute to the classification or prediction of the model [69]. ShaPRap, for instance, underwent testing with non-technical users and findings revealed that some of them struggled to interpret features accurately [69]. However, additional explanations are often provided using natural language [69, 76]. As we focus on domain experts, they may have the necessary knowledge to interpret the features correctly, as they are used to the jargon used in the domain.

Insights

The visualisation mantra *"first overview, zoom and filter, then details-on-demand"* [79] is commonly employed by interfaces. This is not without reason, because it involves important interactions that are important in high-risk environments such as law enforcement. This mantra allowsusers to access raw data at any moment, which can be important evidence in convicting individuals.

Often, an overview of all data elements is presented initially. This can be filtered and zoomed in to access more detailed explanations. Furthermore, various interfaces offer additional descriptions in plain language to ensure that all users can understand particular features.

However, all of the interfaces in the section above have been developed for domains other than the law enforcement domain. LegalVis [75] and VBridge [76] are still somewhat close to our domain in that they also have to support decisions that affect the lives of human beings in high-stake domains. However, VBridge has been developed for the health care domain and LegalVis has been developed for the legal domain. In these domains, there can sometimes be more time for reflection on what is the right decision, because action does not have to be taken immediately. In law enforcement, this is not the case. Decisions are often made that must lead to immediate action, or it is too late.

The above tools demonstrate that the methods LIME and SHAP can be visualised in different ways for the user. However, these methods have not yet been visualized within the law enforcement domain. Therefore, the insights gained in this literature review serve as a guide for designing an interface for non-technical experts in the law enforcement domain. Further in this research, through evaluation sessions, we will see if the proposed design contributes to the decision making process.

From the literature review, we can identify several XAI interface design themes that explain local ML model decisions to users via LIME and SHAP. Each of these themes serves different purposes and employs distinct visualization techniques. We will incorporate these themes into our design, along with the decision makers' needs and requirements that we ascertain through interviews, to visualise local explanations of an ML model for decision makers. The themes are:

1. **Overview of Data:** provide an overview of all items in the data set. Through a variety of interactions such as filtering, selecting, and ordering, the user is able to find specific and important items first.

2. **Explanations on request:** provide necessary additional information about what various features mean, or what the "normal" score would be for features using natural language.

3. **Provide extra explanation in natural language:** to ensure that the explanation does not depend solely on the user's interpretation, additional information, or labels, should be explained in natural language, if the user needs it.

4. Interviews

The goal of our research is to find out how to effectively visualise local explanations of machine learning (ML) models so that law enforcement decision makers can better understand them and incorporate them into their decision making processes. To answer this central question, we must first answer sub-question "*SQ2: Who are the stakeholders that potentially interact with ML models in the law enforcement domain, and what are their needs related to the decision making process?*". This is also the discover step in the nine-stage framework.

The common types of XAI explanations were discussed in the preceding two chapters. In Chapter 2, it was determined that visualisations are employed to elucidate ML models. However, varies methods exist by which these visualisations can be developed, as discovered in Chapter 3. Subsequently, Chapter 3 analysed existing tools that use visualisation to explain ML models. In this Chapter we found that there has not yet been a tool developed that uses visualisation to explain ML models specifically in the law enforcement context. We noticed that numerous ML tools are in development for technical experts, but there are fewer options for non-technical experts who are domain specialists.

Semi-structured interviews were conducted to gain insight into stakeholders who may interact with ML models in law enforcement. The interviews sought to determine their decision making needs, which information they use in the decision making process, what tools they use, and with whom they communicate. First, we discuss the recruitment process in section 4.1. Then the final set of participants is discussed in section 4.2. Section 4.3 discusses the design of the semi-structured interviews. The analysis of the interviews is discussed in section 4.4. Finally, the termination rule is mentioned in section 4.5.

The Ethics and Privacy Quick Scan of the Research Institute for Information and Computing Sciences of Utrecht University was conducted (see Appendix A). This scan classifies this research as low risk, which indicates that it does not require a more comprehensive ethics review or a privacy assessment.

4.1 Recruitment

This research was conducted in cooperation with the Dutch National Police. The participants in the interviews were only people employed by the Dutch police, from units throughout the Netherlands. Therefore, it was not a selection criterion that all participants should come from the same unit.

Participants were recruited using a combination of snowball [80] and convenience [81] sampling to diversify the participant base, allowing for the inclusion of individuals who may not have been part of the original network. The first three participants were identified via convenience sampling and were acquaintances of

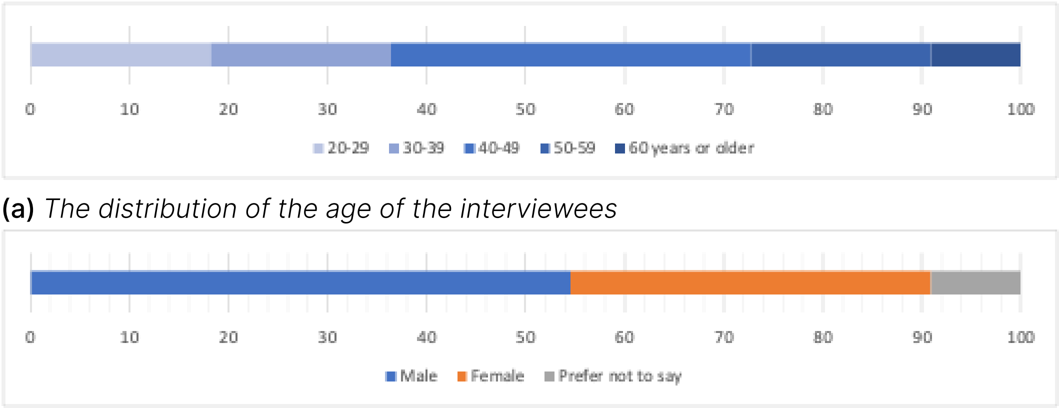
ID	Gender	Role	Years experience	Speciality	Decision	Analysis
P01	F	Data scientist	3	Artificial Intelligence	3	3
P02	M	Data scientist	5	Data science	3	3
P03	M	Data scientist	4	Techinal informatics	2	3
P04	-	Team manager at TROI	-	-	-	-
P05	-	Software developer	20+	IT	4	2
P06	F	UX designer	3	Design/ICT	2	2
P07	F	Product owner	13	Detective science/ international relations	4	2
P08	F	Senior Intelligence	6	*Not applicable*	4	4
P09	M	Police sergeant	8	*Not applicable*	2	4
P10	M	Operational Specialist	20	Management	5	2
P11	M	events/demonstration coordinator	6	Risk management	4	5

Table 4.1: Demographics and other information about the interviewees. The - means that the data is either unknown or someone preferred not to say. Years experience means the number of years of experience in the position for which they were being interviewed. Decision indicates, on a scale of 1 (not at all) to 5 (very often, across multiple departments), the extent to which decision making is seen as a primary task within the function. Analysis is the extent to which data analysis is seen as a primary task. The scale ranges from 1 (not at all), 2 (not my job, however, I do involve the work of the data analysts to make decisions), 3 (do preform data analysis, but not on a regular basis),to 4 (this is my regular task).

the research team, all of whom were data scientists. These three data scientist facilitated the snowball effect by referring eight additional participants from their network during the interviews. Potential participants were contacted via email to schedule study sessions.

4.2 Participants

A total of eleven interviews were conducted. Full details of the eleven participants are given in Table 4.1. Each participant was given a number by the researcher (P01 to P11). Please note that both the roles and the specialisation were self-identified by the participants. Unfortunately, the demographic survey of participant P4 was not saved due to a technical problem. As a result, we can only identify participant P4's role as this was discussed during the interview. Participants' years of experience ranged from 3 to 20 years. Participants ranged in age from 20 to over 60, with the majority being aged between 40 and 49, see Figure 4.1a for complete distribution. Out of the total of eleven participants, four self-identified as female, four as male, and one participant chose not to identify (Figure 4.1b).



(a) *The distribution of the age of the interviewees*

(b) *The distribution of the gender of the interviewees*

Figure 4.1: *Demographics interviewees*

4.3 Interview Design

The interviews took place from April 2023 to July 2023. Whenever feasible, the interviews were conducted in person at the interviewee's workplace in a private room or space. In four cases, we conducted remote interviews using video conferencing software (Microsoft Teams). The interviews were one-on-one sessions of approximately one hour and conducted in Dutch. One interview exceeded this timeframe, lasting 1.5 hours. The range of interview duration varied from 34 to 90 minutes.

Prior to each interview, participants were provided with information on the study's purpose and procedure and were asked for their informed consent whereby all participants gave permission for the interviews to be recorded. These recordings were used for the transcription of the interviews for the later coding process. All interviews were completed prior to the coding process.

The interviews were conducted in a semi-structured manner with guidance from a list of topics, as indicated in Table 4.2. The methodology of using a topic list during semi-structured interviews enables efficient comparison of the results of each interview [82]. Furthermore, semi-structured interviews have an added advantage of providing the researcher with the opportunity to ask related questions about topics that come up during the interview and are relevant to the research [82]. In addition, it gives participants the opportunity to elaborate on important related issues that may be relevant to the research [82].

Each interview was structured in the same way. After giving informed consent, participants were asked if they could give a brief explanation of what their job entailed, to put them at ease. They were subsequently asked to recall and describe a specific instance in which they faced a challenging decision they had to make within their function. Further questioning was conducted based on Table 4.2 in order to cover all relevant topics.

The complete interview protocol, including the full list of questions, are included in Appendix D.

4.4 Analysis

All interviews were transcribed and analysed in Dutch. For automatic transcription, the MP4 files were imported into Microsoft Word Online. This tool is safe to

1.	Who are the stakeholders that can potentially interact with an ML model in the law enforcement domain?
2.	What types of decisions must be made in the law enforcement domain when utilising ML models?
3.	What types of data do stakeholders in the law enforcement domain use when interacting with ML models?
4.	What technology and tools do these stakeholders use to guide them in their decision making process?
5.	How do these stakeholders communicate their information with others?

Table 4.2: *Interview topic list*

use and is supported by Utrecht University, as the data will not be used or sold for other commercial activities. These automated transcripts formed the basis of the final transcripts, which were subsequently updated manually to ensure accurate alignment with the recorded audio. The transcripts and audio files were stored confidentially on a secure server at Utrecht University. Upon completion of the investigation, the audio files will be expeditiously deleted.

Certain quotes have been translated as precisely as possible from Dutch to English for inclusion in this research. Grammatical modifications have been made in a few quotes to ensure comprehensibility.

The NVivo analysis software tool was used to code the data [83]. Selective coding [84] was employed to code the data. In order to do this, categories were established, these being: *stakeholders*; *data*; *tools and techniques*; *decisions*; and *communication*. The classification process involved selecting and assigning transcribed passages to the relevant categories.

4.5 Termination

The iterative snowballing process continued until almost no new insights were found. After conducting 11 interviews, we stopped data collection when we reached theoretical saturation.

In the following chapter, the outcomes of the interviews will be discussed.

5. Interviews Results

In this chapter, we will discuss the findings from the interviews. These findings will be structured using the topic list from Table 4.2: the different stakeholders who (potentially) interact with ML models (Paragraph 5.1.1); the decisions that need to be made when utilising ML models (Paragraph 5.1.2); the data used when interacting with ML models (Paragraph 5.1.3); the tools and techniques used to guide decision making processes (Paragraph 5.1.4); and the way relevant information for decisions is communicated (Section 5.1.5). Finally, one specific case study is presented to empower the focus of the study (Section 5.2).

5.1 General Results

First, the general results of the interviews will be discussed, without focus on any particular type of ML model. In Section 5.2, the results of the interviews which did focus on a specific model will be presented.

5.1.1 Stakeholders Interacting with ML Models

Different stakeholders interact with an ML model in the law enforcement domain. This interaction takes place among data scientists during the development of the model and possibly during an iteration to adjust or improve the model.

For other roles (Team maneger, software developer, UX designer, product owner), this interaction takes place when the model is already developed by the data scientist. These roles ensure that the model is integrated in the desired way in its own environment. What the desired way is depends on which model it is, and the environment in which it is to be used. For example, a model can be used by police officers on the street, then the officers have to interact with it on their phones. However, a model can also be used in the office, in which someone should be able to interact with it on a computer or laptop.

Finally, end users interact with the model. This interaction only takes place after the model has been developed and integrated into the system in which it will be used.

Before any interaction with a model can take place, regardless of the role, there must be a client asking for a model. A problem that was reported by all three data scientists is that they often develop the model proposals themselves. However, this is not ideal and has the consequence that they do not know what information is relevant to the end user, as participant P02 noted:

> *"Actually, the police officer does not really know what to ask the data scientist. And we [data scientists] also find it very difficult to create something for him that he can use in practice. So the hardest thing is always to get the right question out of the client. ... But ideally you do not want that [come up with an idea for a model yourself] to be the case, because then you often decide for someone what he or she needs, without actually understanding what he or she needs."* - P02 Data scientist

The issue lies in the fact that data scientists hold expertise in developing model techniques and methods, but lack the necessary domain knowledge to determine which information is crucial for the decision making process. In an optimal scenario, the request would originate from within the organisation, with an end user specifying the required and the relevant information. However, participants P01 and P07 revealed that end users often have no idea what to ask or expect from the data scientists:

> *"... people often have a problem, but then someone else has to signal that an AI solution can be made for it."* - P01 Data scientist

> *"But we also struggle a bit with the fact that end users [The individuals who utilise the outputs of the model in decision making processes] also cannot always oversee everything that is possible that can make their work easier. So end user questions are very often of things they already know then."* - P07 Product owner

5.1.2 Decisions Made Utilising ML Models

An important aspect when using ML models is that end users must be aware that ML models can also generate incorrect outputs. This is evident from both the interview with P04 and the interview with P02.

P02 therefore says that users should view a model more as a tool rather than as an automatic task. In addition, it is important that the user understands how the model shows the results. A model may be able to show the top ten best results, but this does not mean that there are no more relevant results after those ten results. For example, if a dataset needs to be searched for people who trade cannabis, the model can provide a list of the most likely people you are looking for. Yet the user needs to understand that there may be more people in the data set who are likely to be dealing cannabis.

Furthermore, it is important that a model is always accessible, that it can be used responsibly by decision makers, and that the model is transparent. This applies not only to the model itself, but also to how the code is written, where the model is active and who has access to the data.

Additionally, it is very important that end users can always see the entire data set. This is because the model may contain errors that cause important information to be disregarded, as noted by participants P07 and P06:

> *"But we also have to give them the possibility to see them, because there could be something in the lower confidence score. So we always start with everything."* - P07 Product owner

> *"And it is always possible to see everything, so we do not have some kind of cut or point of if it is below 0.2 or so."* - P06 UX designer

Lastly, it is necessary that the end user can always go back to the forensic source of the information, because information generated by a model cannot be used in an investigation. For example, if an interview is automatically transcribed, the forensic source is the audio recording.

5.1.3 Data Used Interacting with ML Models

Availability of sufficient data is a crucial factor in model development. However, the data scientists themselves are not allowed to access this data themselves and are dependent on the person who wants to apply the model in practice (the client). Yet, acquiring the data is not always easy for data scientists, as P01 noted:

> *"It [data] is often stored in different ways by various owners, making it challenging to locate individuals who agree to utilise it."* - P01 Data scientist

The exact type of data that is used depends on the purpose for which the model is being developed. Several models were discussed during the interviews. Each model had a different purpose. For example, one model under discussion was a model which would assist in the prioritisation of documents. This model would classify the reports (written by the employees) according to topics. For example, a report might contain information about weapons or drugs. Investigators could then use this model to search the reports if they were looking for documents that might be relevant to a case involving someone suspected of dealing drugs.

Another model that was discussed is one that helps to identify licence plates when someone is driving and talking on the phone. The model will then pass on the images that it thinks show that someone is talking on the phone while they are driving. A human then has to judge the image and fine may or may not be imposed.

Another example is a model that converts speech into text. This model is used in interrogations. It transcribes the interrogation live so that it can be stored in a report. This report can then be used for further investigation/trial.

As P06 emphasised, the model's purpose depends on the operation's needs:

> *"it very much depends on the demand from the operation "* - P06

Nonetheless, the provision of accessible and meaningful information is critical as it is essential that end-users (the decision makers) have access to relevant information to facilitate their decision making processes. However, experience shows that more information is not always better. This was highlighted by participant P02, who noted:

> *"A mistake we used to make a lot is giving way too much extra data on top of the classification itself. ... For the customer, it's just noise."* - P02 Data scientist

In addition, P01 and P06 reported that scores are often difficult for end-users to interpret, even though they are often provided as information (or at least the accuracy score).

> *"Sometimes we also say what precision and recall are, but then you have to explain them again. And often you find that you lose people then".* - P01 Data scientist

> *"People thought it [accuracy scores] was so unclear that we removed it. So we changed the score to categories: very low, low, medium and high. ... We really deliberately removed the scores because it caused a lot of confusion."* - P06 UX designer

Finally, data quality is also crucial in ensuring model accuracy. If the data collected by street agents is of inadequate quality, the model will not be accurate. For example, in the model where you can search by topic, agent reports are used as data. If agents provide little detail in their reports, it is also difficult for a model to assign a topic to them. This means that decision makers are not only dependent on the information that they get from a model (provided by the data scientists), but also on the way in which the data is stored by their colleagues.

Suppose a colleague arrests someone suspected of dealing drugs. This person may have been stopped during a traffic control and a large quantity of cocaine may have been found in his trunk. If an officer simply wrote down that someone suspected of dealing drugs was stopped, but did not mention what was found or how the person was stopped, the decision maker would not have enough information to convict the person.

5.1.4 Tools and Techniques that Guide Decision Making Processes

Again, it depends very much on the problem at hand which tools and techniques are used by decision makers. Sometimes decision makers need to search through very large amounts of data to find the relevant information. Other times, they just need to check that the person on the image is indeed on their phone while driving. Therefore it depends very much on the context.

Yet, interfaces are always designed in the same way to make it easy for end users (decision makers). For example, buttons have the same appearance and are often in the same position. This means that the end user is already familiar with the design and only needs to get to know the new model.

5.1.5 Communication of Relevant Information

In practice, it is common for data scientists to develop a model and deliver it to the customer, without receiving ample feedback about the workings of the model. For example, they receive little feedback from the client on whether or not something is missing from the model.

This problem is evidenced by the interview with P01:

> *"I have also sometimes set up interviews with users to ask them what they think of the way the model works, but it is quite difficult to get that information from users. ... It is often the case that if it [the model] works, it gets used and you don not hear about it, but if it does not work, you often don not hear about it either, because they just stop using it."* - P01 Data scientist

However, it seems that communication is happening between the end users and the UX designers. The UX designer then schedules interviews or sessions to ask questions about what the user is up against, what their work is now, and then they start looking at how to turn this into a design. However, these designs will only be discussed with the team, and then the developer will start to work on them. This information isn't passed on to the data scientists. This is a problem because the information provided by the model remains the same, it may just be presented differently.

Summary

To sum up, various roles interact with ML models. However, each of these roles has different goals when interacting with the ML models. This is consistent with the findings in the literature discussed in Subsection 2.3.1. Nonetheless, the issue in the police is that decision makers, who are the end-users, do not know the specific goals for which an ML model can be developed, leading to data scientists being unaware of the end-users' needs and requirements.

In addition, there is very little contact between the data scientists and the end users (the decision makers). Yet, there is contact between the team responsible for integrating a model into the environment in which it will be used and the end user. Nevertheless, the problem of insufficient information in the model output remains.

Furthermore, the use of a model should be carried out in a responsible manner, with careful consideration given to the use of data and authorised access, and by explicitly stating that models have the capacity to make errors. Further, it is important that the complete data set is visible to the end user at all times.

We have seen several examples of models that each have a different purpose and are used in a different context. The type of data used also differs. One of them used text (topic classification), the other one used images (number plate recognition) and another one used audio files (speech to text). To delimit our research, we have chosen to focus on a specific context. The last three interviews are therefore focused on this specific case. We will now explain this case in the next section.

5.2 Specific Case study

To limit the scope of our research, we decided to focus on one specific case study. We then focused on this case study in the last three interviews to understand the needs of the decision makers, so that we could use these insights to design local explanations of ML models for these decision makers.

In this Section we will first discuss the context of this case study (Paragraph 5.2.1). Next, we will discuss the data that the decision makers within this case study use in their decision making processes (Paragraph 5.2.2). Finally, this Section outlines decision makers needs and requirements (Paragraph 5.2.3). These insights will inform our design process.

5.2.1 Decisions Made in Case Study

In this case study we will focus on the field of public order and safety. The purpose of the police in the field of public order and safety is to ensure that public order is maintained and that everyone who is in public order is safe. To ensure safety, the police will take action against anyone who poses a threat to the public order. This may mean letting you off by warning or fining you without using force, but it may also mean arresting you.

The aim of the decision makers, within this field, is therefore to determine which units should be deployed, how they should be deployed, where they should be deployed and when they should be deployed. For instance, it is possible that a particular incident will require the attention of a local policeman, but it is also possible that a mobile unit will be ready to intervene if an incident gets out of control.

In order to make such a decision, it is important that the decision makers have sufficient information about an upcoming incident. We will refer to incidents here as the general category. Within this category, events, demonstrations, public order permit events, and paid football events can take place. With this information, they can then formulate an action plan and determine the deployment for this incident. Figure 5.1 shows the steps involved to acquire the information.

5.2.2 Data Used to Guide the Decisions

In order to decide whether an incident poses a problem for public order and safety, a signal must first be given. Without a signal, the police do not know that an incident will take place. A signal may be received because the municipality has received an inquiry about an incident. Yet it is also possible that a signal arrives at the police station or through the local police officer.

The next step is to analyse this signal. For example, the police will check who made the signal, whether there has been a previous incident with the same organisers and how that went. This information is important because, for example, someone may have made a signal about an incident but nothing will

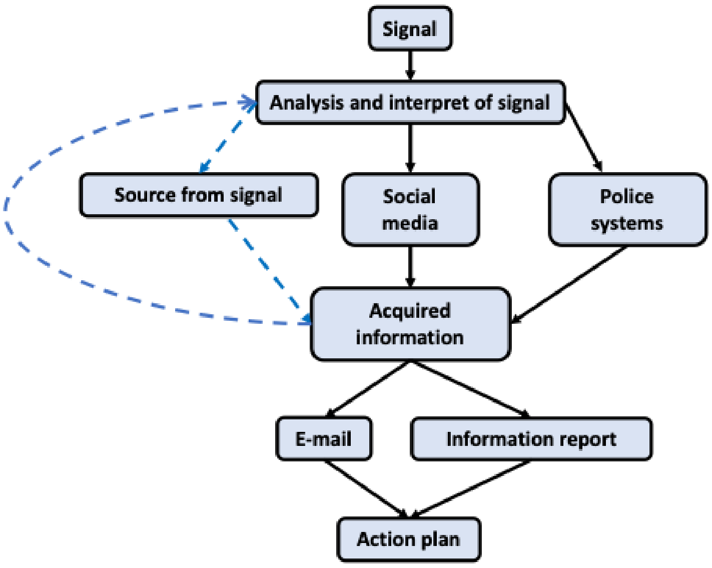


Figure 5.1: The summary of the actions taken to acquire the relevant information of the possibility of disruption of public order by a signal. The blue dotted arrows show a possible option, which is not always taken. The black solid arrows are always executed.

actually happen. Besides, it might be that an organisation has been uncooperative with the police at other times and has frequently been a threat to public order.

The police also checks if there is a counter-event. For example, during a demonstration, a counter-demonstration may be planned, which could result in both parties facing each other at some point, which could cause friction. This could mean that more police resources are needed.

In addition, information is also gathered on the expected number of participants, the location of the incident, whether it is a static or dynamic incident, the target audience, the time of the incident. For example, if a large number of visitors are expected, and they are all over the age of 18, it is possible that alcohol will be consumed. If the event is dynamic they may pass through places where there is a difference of opinion, which can lead to friction. As a result, some combinations may require more resources, while other combinations may not pose as much of a threat to public order.

The information is collected through police systems and social media. If an incident took place earlier, the police systems can provide information on whether any arrests were made, what units were deployed and what the atmosphere was like during the incident. Social media is mainly used to find out how many people are expected to participate. To give an example of how this is done: the police look at how many people have signed up for a Facebook event. Social media is also used to see how an incident is shared, to get an idea of how often a message is shared and whether there are people who have a negative opinion about the incident.

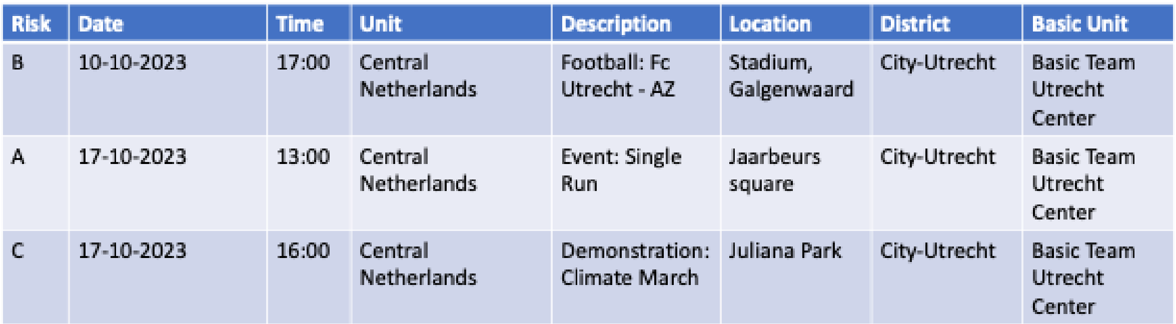
Once the information gathering is complete, the results are compiled into a document that is disseminated among decision makers. Due to the large number of police systems, the document is usually sent by e-mail to ensure that the intended users receive the information.

The decision maker uses the information provided to develop an action plan that includes multiple scenarios. These scenarios outline the expected responses by units to any incident, from the most favourable to the most extreme scenario. For instance, whenever an incident occurs outdoors, there is a predefined protocol in place for dealing with extreme weather conditions. Additionally, there may be a scenario for dealing with a confrontation between two opposing parties, for example during a demonstration.

Such an action plan describes which, how, and when units will be deployed. For instance, a local police officer might only be deployed for visibility purposes. Similarly, a mobile unit may be used to arrest people. Additionally, evacuation procedures for a square or nearby region are being considered, as it is inadequate and unsafe to send people into one direction. For example, if a protest occurs at the outskirts of a city, it would not be advisable to relocate all the attendees within the city, as there may already be people shopping in the city. Sending more people in that direction would make the place even more crowded and unsafe for the people already there.

An action plan with relevant scenarios is translated into a script to ensure coordination between all units during the operation. With the action plan, officers are working towards a common goal rather than working independently of each other. Unfortunately, no access has been provided to such an example.

Furthermore, all incidents are also documented in a Police Calendar. This ensures that everyone is aware of the incidents that will take place in the district. The location, the units responsible and a brief description are included in this calendar. This information can be seen for incidents in the upcoming week, for example, but also for the month and the year. Incidents can also be viewed for previous weeks, months and years. In order to access further information, the user can select an individual incident. Figure 5.2 shows a simplified, anonymous police calendar, which provides a visualisation of incidents. Although it does not contain any specific examples, it shows how incidents are presented based on a list.



Risk	Date	Time	Unit	Description	Location	District	Basic Unit
B	10-10-2023	17:00	Central Netherlands	Football: Fc Utrecht - AZ	Stadium, Galgenwaard	City-Utrecht	Basic Team Utrecht Center
A	17-10-2023	13:00	Central Netherlands	Event: Single Run	Jaarbeurs square	City-Utrecht	Basic Team Utrecht Center
C	17-10-2023	16:00	Central Netherlands	Demonstration: Climate March	Juliana Park	City-Utrecht	Basic Team Utrecht Center

Figure 5.2: *A simplified visualisation of the current police calendar*

5.2.3 Needs and Requirements

We can now answer sub-question "*SQ2: Who are the stakeholders that potentially interact with ML models in the law enforcement domain, and what are their needs related to the decision making process?*".

There are numerous stakeholders that could interact with ML models in the law enforcement domain, so we have chosen to focus on public order and security stakeholders. Although they do not currently use ML models, they may consider incorporating them in the future. An ML model could, for example, assist in the allocation of resources. Therefore, the main needs for such a model in the decision making process are:

- The source of the notification must be present.
- The identity of the organisation must be established in order to check whether there is any prior knowledge of it.
- The information collected should be shareable to enable it to be distributed effectively.
- There needs to be a means of distinguishing between different types of incident. This is because the specific combination of factors involved, such as the time of day and the target group, will require the police to vary their level of deployment
- To achieve a full understanding of what is about to happen, the key aspects of the incident need to be presented in a way which enables all concerned to see what is going to happen in the coming week, month or year.

We have completed the discovery phase of the nine-stage framework and can now move on to the design phase. The interviews have provided us with valuable insights that we utilised to create two initial low-fidelity designs. We will explore both designs and subsequent iterations in chapter 6.

6. Design

This chapter outlines the iterative procedure for producing prototypes to create a tool that visualizes the local explanations of an ML model to public order and safety decision makers. Firstly, section 6.1 will explore the utilisation of interview findings and literature study outcomes in building low-fidelity prototypes. Secondly, we will discuss the iterative process through which the low-fidelity prototypes were adapted until they were finally translated into a high-fidelity prototype in section 6.2. Finally, section 6.3 will discuss in detail the final design used in the evaluation sessions.

6.1 Requirements Design

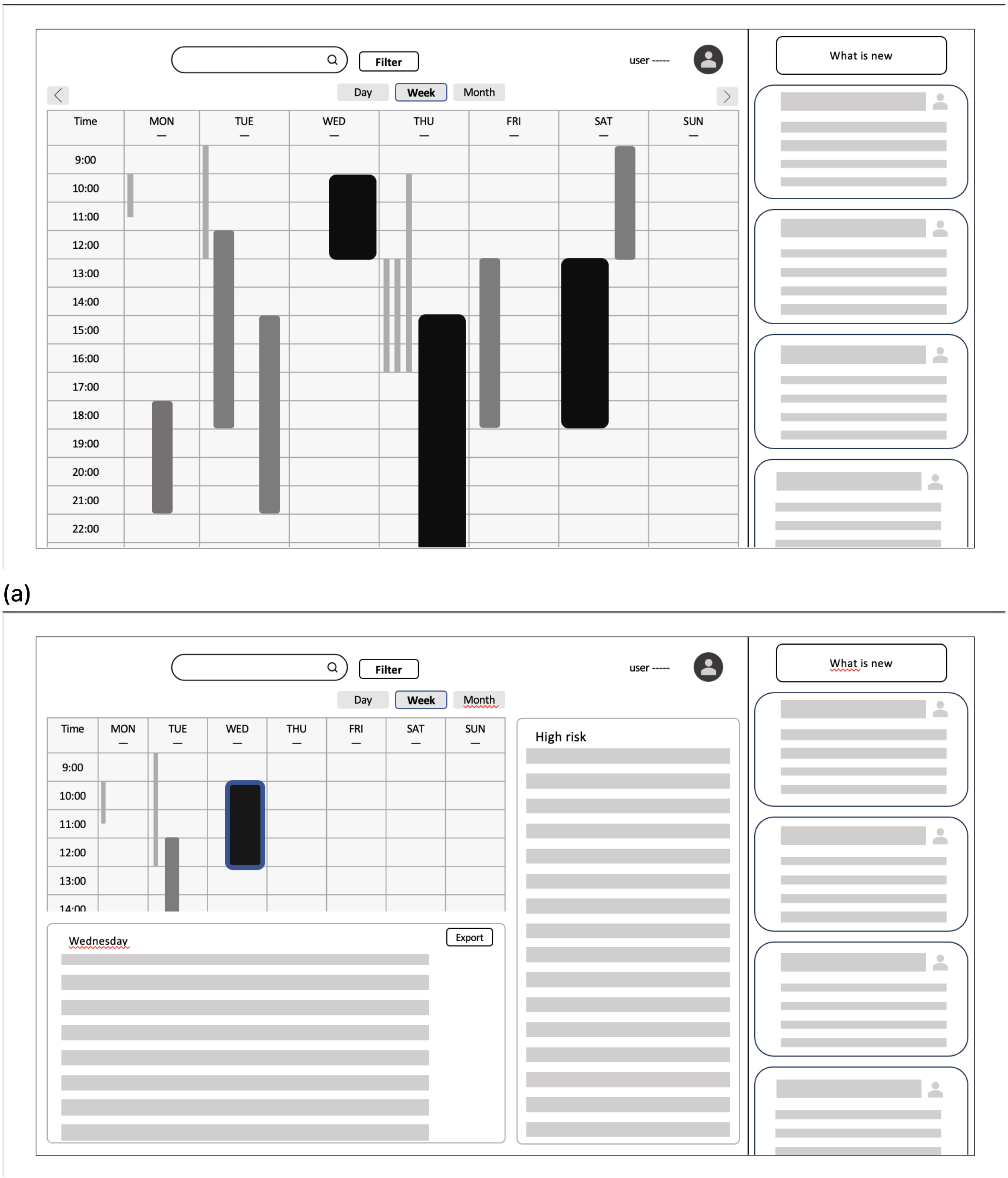
The following requirements for a user interface have emerged from the interviews and related work:

- The source of the notification must be present.
- The identity of the organisation must be established in order to check whether there is any prior knowledge of it.
- The information collected should be shareable to enable it to be distributed effectively.
- There needs to be a means of distinguishing between different types of incident. This is because the specific combination of factors involved, such as the time of day and the target group, will require the police to vary their level of deployment
- To achieve a full understanding of what is about to happen, the key aspects of the incident need to be presented in a way which enables all concerned to see what is going to happen in the coming week, month or year.
- To provide extra explanations for the local explanations of ML models in natural language.
- To provide explanations where necessary. In other words, by making all the necessary additional information visible at some point.
- To provide an overview of the full dataset.

6.2 Low-fidelity Prototypes

Ultimately, we decided that the current Police Calendar needed to be used as the basis for our design. We did this because (1) decision makers already use this tool to make their decisions and are therefore partly used to the way the information is presented, and (2) it partly shows the most important information. However, the incidents are now presented in a list-based format. This list visualisation makes comparison of incidents difficult and the different types of incidents are not immediately visible. In addition, the risk of an incident is indicated by a letter, which could be seen as a shape channel. However, there is evidence from the literature [62] that this is the least effective channel.

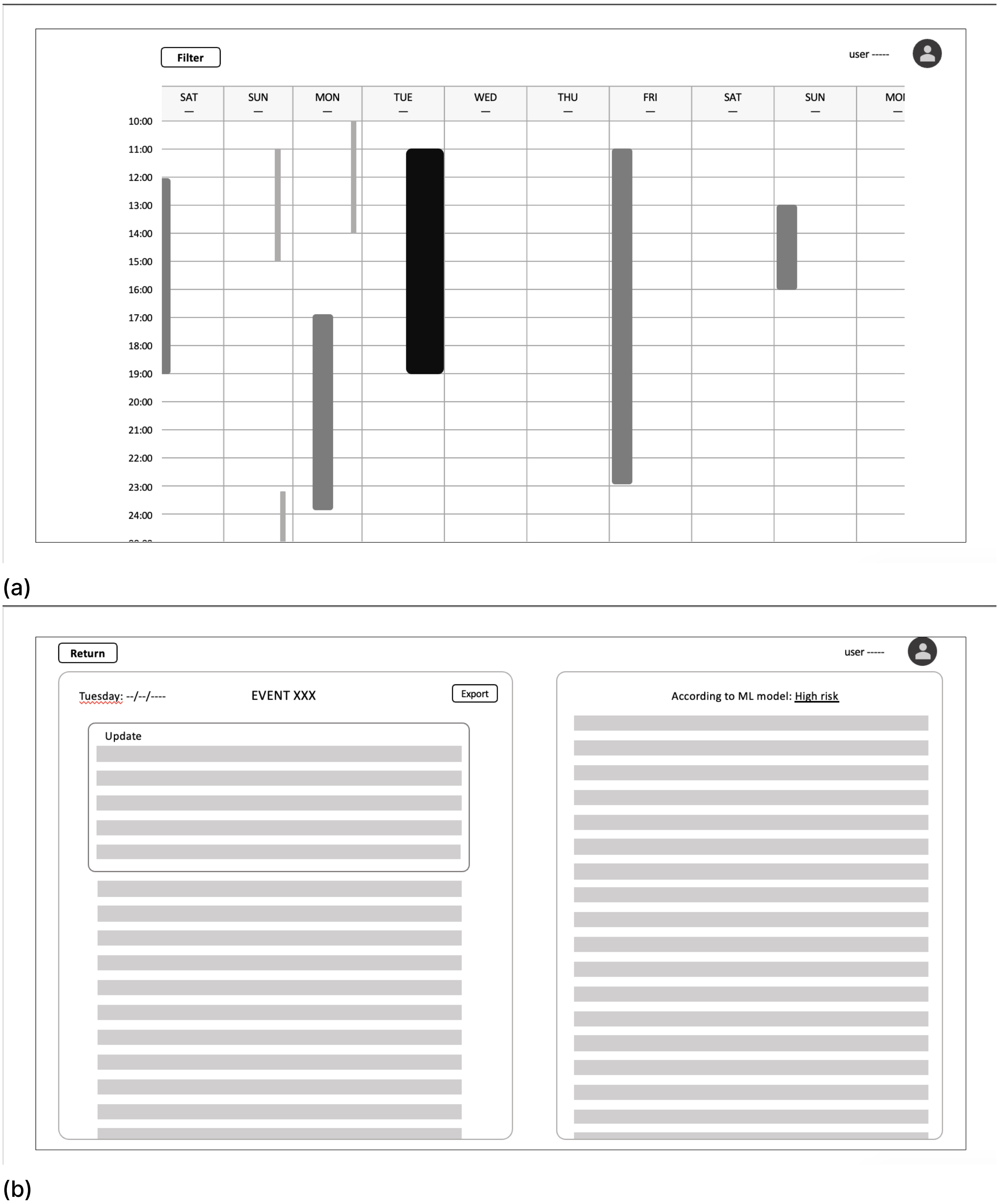
These findings led to two initial low-fidelity prototypes, (see Figures 6.1 and 6.2). These prototypes were created using PowerPoint software. This approach allows for easy adaptation in future iterations. There was no use of colour in the



(a)

(b)

__Figure 6.1:__ (a) Home screen of the first design. (b) Second page after selecting an event, that would give more information



(a)

(b)

Figure 6.2: *(a) Home screen of the second design. (b) Second page after selecting an event, that would give more information*

initial designs as this is not the most important channel according to Munzner [62]. A number of important design decisions were made in the creation of these low-fidelity prototypes, and these will be discussed below.

6.2.1 Scope of the design

First of all, we decided to narrow down the scope of the research a little bit more. For example, during the interviews we only spoke to decision makers in the district of Stad-Utrecht (Utrecht City). For this reason, we have decided to base the design only on incidents that are going to take place in this district. We also decided to focus only on incidents that will take place in October, so that we can ask specific questions about them during the evaluations.

Finally, we decided to use only incidents from the following four categories: events, demonstrations, football events and public order events. There are several other categories that the police distinguish, but the interviews showed that these are the most common categories.

6.2.2 First Two Designs

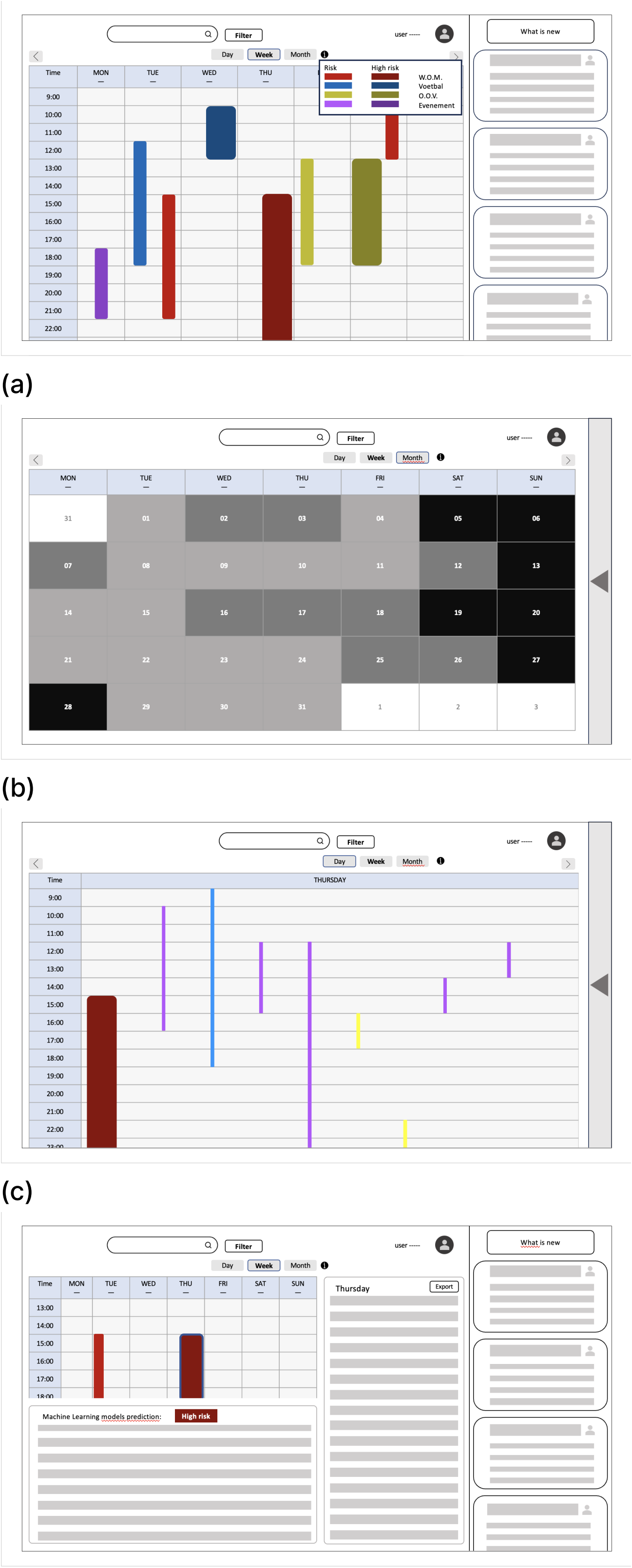
The most important information we wanted to convey was the expected number of resources per incident. This is determined by both an ML model and a human. Munzner's [62] ranking shows that colour is not the most effective channel when other channels are not being used. Therefore, we decided to leave colour out of the initial designs.

Since we wanted to ensure that people could see at a glance when an incident occurred, we decided to use a calendar view, see Figures 6.1 and 6.2. Here the incidents are shown by means of a bar. As a result, the encoding for the location and the length are already in use. The length indicates the duration of an incident. We then decided to use both the size and the luminance for the indication of the risk. In a later iteration we realised we were using two encodings and changed this, but in these designs we used both encodings.

As you can see, there are three different levels of risks: *no risk*, *some risk* and *high risk*. The lighter the colour of the bar and the narrower the bar, the lower the risk. A high-risk incident would require a significant police deployment, while a low-risk situation necessitates police deployment, although not yet on a large scale. No risk means that no police deployment is to be expected. Even though there is no risk to the police, it is important that the incidents are reported, as this can sometimes influence the commitment to other incidents.

Discussions were held with both supervisors and domain experts. The result was that the initial designs were very similar but already in line with the users' needs. Two feedback sessions were held with two different domain experts. The results showed that the domain experts understood the visualisation well enough to identify a high-risk incident. The initial designs also did not include a legend explaining what the colours and thickness meant. This was also not given to the domain experts in advance. However, they were able to correctly interpret this information from the visualisation.

An important point that emerged from the feedback sessions was the need to distinguish between the different incidents. This is important in order to effectively determine the deployment required by the police. In addition, the first design was preferred because it showed all the information about an incident without losing the ability to see the calendar. This is an important difference from the current version of the Police Calendar, which also requires switching screens to read sufficient information, as similair to the second design.



(a)

(b)

(c)

(d)

Figure 6.3: *(a) Home screen of the third design. This shows the incidents that have a risk and a high risk within the current week. (b) Second page after selecting the month button, this will show the indication of the risk for the police on a day. Darker the colour means higher risk. (c) This shows which incidents take place on the current day. It shows all incidents no matter the risk. (d) This screens shows the information that one will see after selecting an incident.*

In the next iteration it was decided to add more detail. See Figure 6.3. The next iteration was a continuation of the first design.

6.2.3 Third Design

To differentiate between the different incidents, we have added colours in this design. Here, the darker the colour, the higher the level of risk. In addition, we have chosen to include only those incidents that are high risk in the visualisation of the colour of the month. This should make it easier to see which days are likely to require a lot of effort. When the overview per week is shown, only risk or high risk incidents are visible. However, if a specific day is selected, all incidents are shown, even those that do not pose a risk.

We have again chosen to use both encodings for the risk, so both the colour and the thickness. The choice of colours has been a very difficult task. We were also unsure whether to use colour only to indicate risk or to use colour for both risk and incident category. In the end we chose the latter.

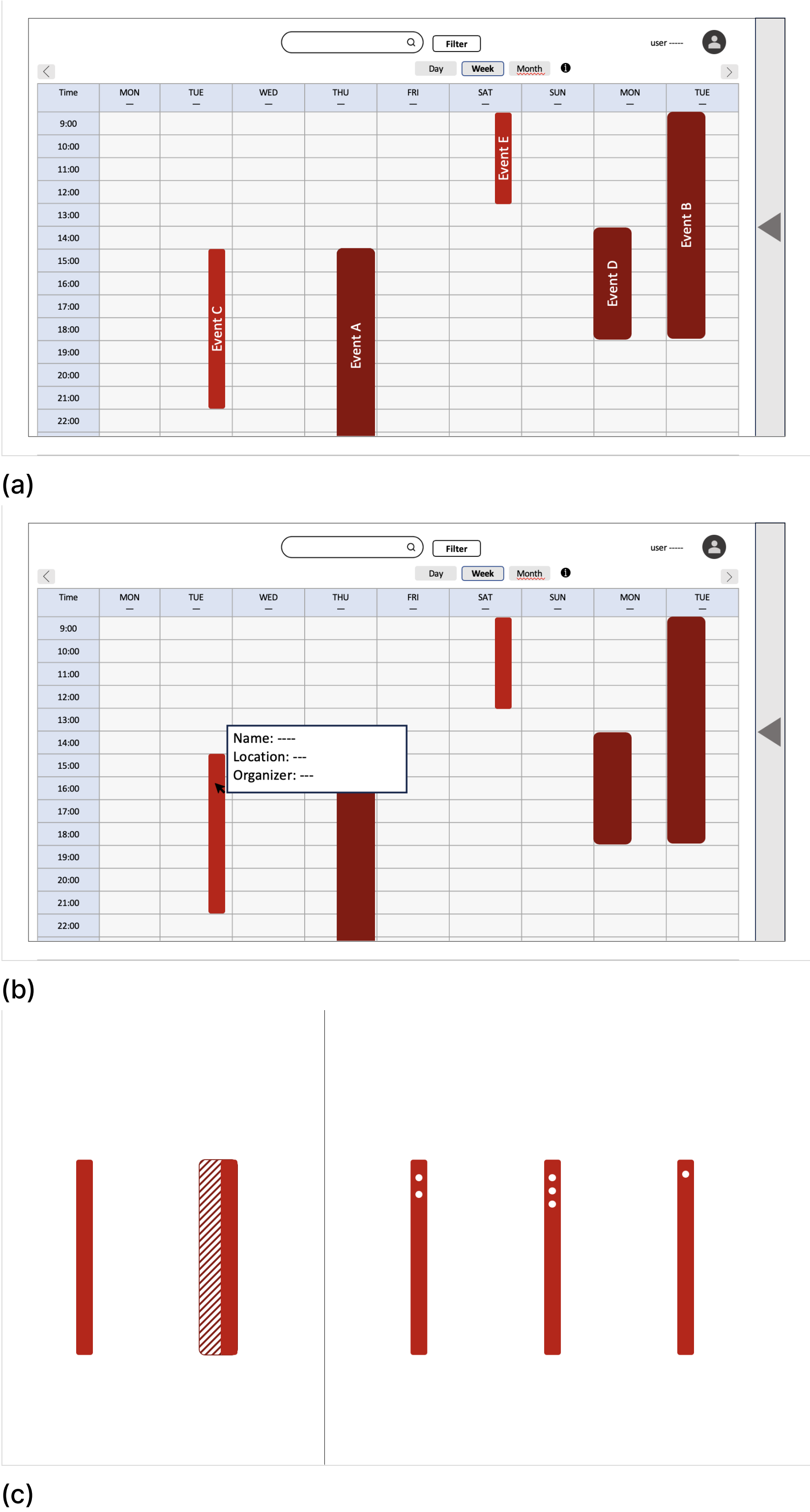
Finally, we made some small designs to show some more detail. We discussed these with the domain experts at the same time as the third design. First of all, it is important to be able to quickly see some information about the incidents. Not just the category it falls into, but also information about the organisation and the location. We had two designs for this, see Figure 6.4. In the end, we decided to use design 6.4b in our high-fidelity prototype. This is because it shows the most information in a structured way.

We also discussed the two designs that would have to demonstrate the classification of the human and the classification of the model. In the left design in Figure 6.4c, the right side shows the models classification and the left side shows the human's classification. If the human has not yet given a classification, you will only see one bar. The stripes indicate the human's classification. The more stripes, the less confident the person is. The pattern on the right shows the classification of people by the circles. The more circles, the more insecure the person. In our final design, we ultimately decided to use the stripes. The classification of the model is indicated by its colour. The darker, the more certain the model is.

The third design and the insights from the feedback session formed the basis of our high-fidelity prototype.

6.3 Final design

An important adjustment to the final design compared to the third design is that we now only use the colours for risk classification, rather than the thickness of the bars. In addition, in this final design we have also added the visualisations for the local explanation of the ML models. These explanations are given by highlighting the feature (which in this case is) in the text. Then, for each feature, it is indicated below how much it contributes to the total score. The total score ensures that an incident is classified and therefore given a risk rating. An explanation is always given. This explanation indicates which total score is low risk, which is high risk and which is high risk.



(a)

Name: ----
Location: ---
Organizer: ---

(b)

(c)

Figure 6.4: (a) First design of showing more information of the incident,(b) Second design of showing more information of the incident, (c) Designs for the risk classification for both the model and human.

This time we also made a conscious choice for the different colours used for the incident categories. We carefully selected a colour palette for qualitative data with four values using ColorBrewer2 [85], with the aim of ensuring that it is suitable for printing and accessible to those with colour blindness. Plans of action based on the acquired information are formulated and disseminated to officers on duty in the police force. In certain instances, these plans are printed, so this was the main feature of the colour palette. Applying these criteria, we arrived at the following palette:

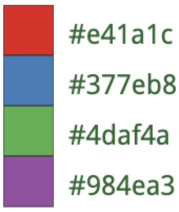


Figure 6.5: *Colour palette*

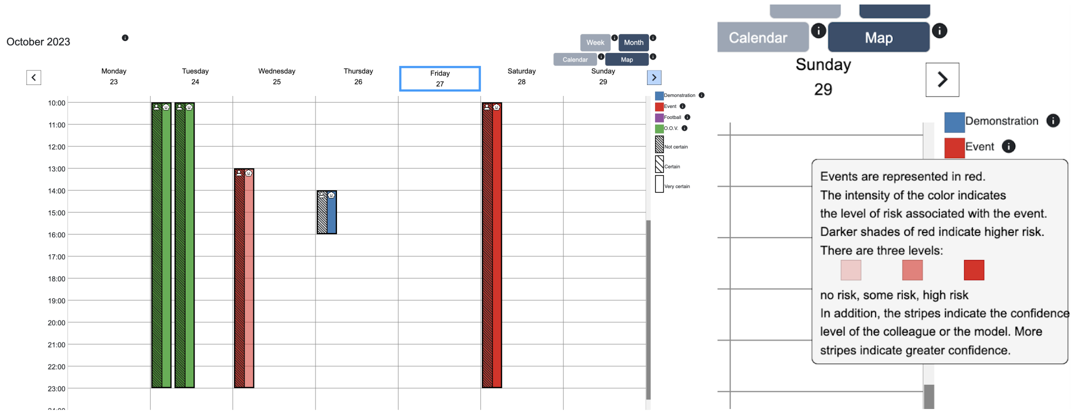
6.3.1 Home screen

Upon entering the tool, users are presented with the home screen shown in Figure 6.6a. The user sees an overview of all incidents occurring during the current week. A legend is displayed on the right side of the screen, indicating which colour represents which type of incident. This legend is always present. When the user hovers the mouse over one of the colours in the legend, it is also shown that there are three different risk classifications and that these are indicated by the darkness of the colour, see Figure 6.6b.

6.3.2 Information screen

The information display presents data used by both human and model to classify the risk. When users select an incident, they access this information. If they subsequently select the calendar again, the home screen reappears.

The left side displays the data that the human used for the classification, while the right side shows the model's data. The model also explains which features (which are words) contribute to the total score. The colour of the category has

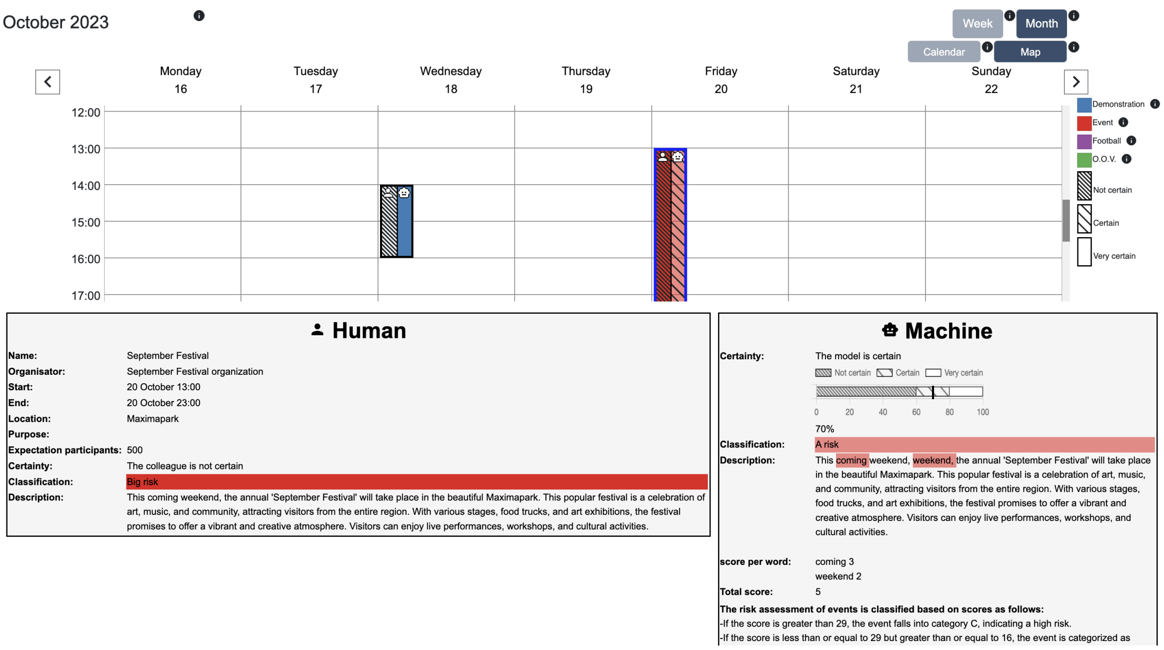


Events are represented in red.
The intensity of the color indicates
the level of risk associated with the event.
Darker shades of red indicate higher risk.
There are three levels:
no risk, some risk, high risk
In addition, the stripes indicate the confidence
level of the colleague or the model. More
stripes indicate greater confidence.

(a) *Screenshots of home screen of the final design*

(b) *The additional information in the legend.*

Figure 6.6: *Final design of the tool*



Human

Name: September Festival
Organisator: September Festival organization
Start: 20 October 13:00
End: 20 October 23:00
Location: Maximapark
Purpose:
Expectation participants: 500
Certainty: The colleague is not certain
Classification: Big risk
Description: This coming weekend, the annual 'September Festival' will take place in the beautiful Maximapark. This popular festival is a celebration of art, music, and community, attracting visitors from the entire region. With various stages, food trucks, and art exhibitions, the festival promises to offer a vibrant and creative atmosphere. Visitors can enjoy live performances, workshops, and cultural activities.

Machine

Certainty: The model is certain

70%
Classification: A risk
Description: This coming weekend, weekend, the annual 'September Festival' will take place in the beautiful Maximapark. This popular festival is a celebration of art, music, and community, attracting visitors from the entire region. With various stages, food trucks, and art exhibitions, the festival promises to offer a vibrant and creative atmosphere. Visitors can enjoy live performances, workshops, and cultural activities.

score per word: coming 3
weekend 2
Total score: 5
The risk assessment of events is classified based on scores as follows:
-If the score is greater than 29, the event falls into category C, indicating a high risk.
-If the score is less than or equal to 29 but greater than or equal to 16, the event is categorized as

Figure 6.7: *The information available about an incident. On the left you see the information used by humans to classify the risk, and on the right the information used by the model. If the person has not yet given a rating, a note appears: The colleague has not yet given a rating. The bar in the calendar turns light grey.*

been reused and its darkness indicates the contribution of a feature to the total score. If a feature has a high score, it will also have a dark colour. The features are scored between 1 and 5.

6.3.3 Month screen

The month screen shows all the incidents that occurred during October, regardless of how these incidents have been classified on the risk spectrum.

Hovering over any incident will provide users with more details. This feature

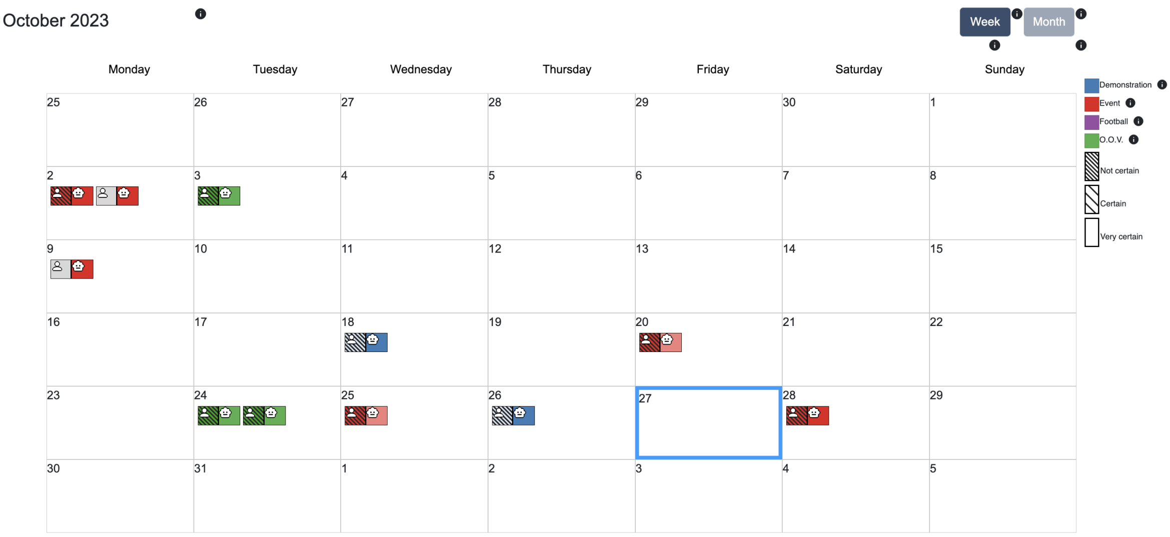


Figure 6.8: *Screenshot of the month screen*

is equivalent to the weekly calendar's incident hover function. Clicking on an incident in either the monthly screen or weekly calendar will provide identical information as well. To return to the month screen, the user can simply click on the calendar again.

6.3.4 Map screen

Finally, the user can see where an incident occurs in the Stad-Utrecht district by pressing the map button. This will only display the incidents that occur per day. Note that this differs from the weekly calendar.

Furthermore, the map button is only accessible on the weekly calendar. The map button allows the user to determine the location of incidents in the Stad-Utrecht district on a daily basis. Once the user clicks on the button, hovering over a pin will reveal the incident's name, while clicking on it once more will provide additional information. This information is equivalent to what appears when you click on an incident in the weekly or monthly calendar.

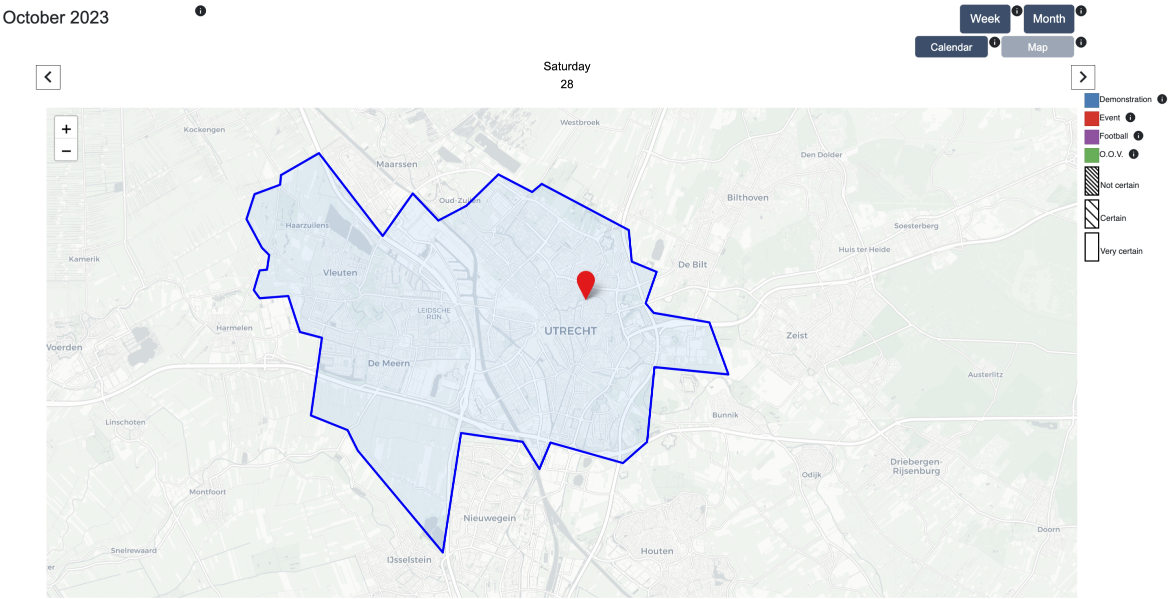


Figure 6.9: Screenshot of the map screen

7. Evaluation

This section describes the evaluation we used to understand the positive aspects and potential areas for improvement of our tool. First, in Chapter 7.2 we discuss the recruitment process. Then the design of the evaluation sessions is discussed in section 7.3. Section 7.4 discusses the analysis of the evaluation sessions. Finally, section 8 discusses the findings of the evaluation sessions.

The Ethics and Privacy Quick Scan of the Research Institute for Information and Computing Sciences of Utrecht University was conducted (see Appendix A). This scan classifies this research as low risk, which indicates that it does not require a more comprehensive ethics review or a privacy assessment.

7.1 Recruitment

This research was conducted in cooperation with the Dutch National Police. The participants in the evaluation sessions were only people employed by the Dutch police, from the unit Midden Nederland (Central Netherlands).

Participants were recruited using convenience [81] sampling. Namely, we asked the three participants who work within the public order and safety domain of the police and who have also participated in the interviews before. This resulted in 3 participants.

7.2 Participants

A total of three evaluation sessions were conducted. Full details of the three participants are given in Table 7.1. Each participant was given a number by the researcher (P01 to P03). Please note that both the roles and the specialisation were self-identified by the participants. Participants' years of experience ranged from 6 to 20 years.

7.3 Evaluation Design

The evaluation sessions were conducted in October 2023. All three sessions were held in-person at the interviewee's workplace in a private room or space,

ID	Gender	Role	Years experience	Speciality	Decision	Analysis
P01	M	Events/demonstration coordinator	6	Risk management	4	5
P02	F	Senior Intelligence	6	*Not applicable*	4	4
P03	M	Operational Specialist	20	Management	5	2

Table 7.1: Demographics and other information about the interviewees. The - means that the data is either unknown or someone preferred not to say. Years experience means the number of years of experience in the position for which they were being interviewed. Decision indicates, on a scale of 1 (not at all) to 5 (very often, across multiple departments), the extent to which decision making is seen as a primary task within the function. Analysis is the extent to which data analysis is seen as a primary task. The scale ranges from 1 (not at all) to 4 (this is my regular task).

and lasted one hour. In addition, all three evaluation sessions were conducted in Dutch, as this is the native language of both the researcher and the participants.

Prior to each evaluation session, participants were provided with information on the study's purpose and procedure and asked for their informed consent whereby all participants gave permission for the evaluation session to be audio and screen recorded. These recordings were then used for the transcription of the evaluation sessions for the later coding process. All evaluation sessions were completed prior to the coding process.

Once informed consent had been obtained, the evaluation sessions consisted of five steps. First, a brief description of the tool was given, followed by a short training session. A detailed outline of the training session can be found in Appendix F. During the training session, stopping rules were applied to some of the tasks, requiring participants to complete them correctly two or three times before moving on to the next task.

The next step in the session was the judgment step. This step determined participants' comprehension of the visualisation through a series of tasks, focusing on their ability to locate relevant information. Their comprehension was further tested by performing additional tasks. They were asked to verbalise their thought process.

Following this, a new data set was presented, and participants were asked to complete a decision making task while, again, verbalise their thought process.

Subsequently, an interview was conducted to identify the positive aspects and potential areas for improvement. Finally, participants were asked if they could complete a short demographic survey.

The complete evaluation protocol, including the full list of tasks and interview, can be found in Appendix F. The questionnaire is the same as we used in the interviews, see Appendix C.

7.4 Analysis

All evaluation sessions were transcribed and analysed in Dutch. For automatic transcription, the MP4 files were imported into Microsoft Word Online. This tool is safe to use and is supported by Utrecht University, as the data will not be used or sold for other commercial activities. These automated transcripts formed the basis of the final transcripts, which were subsequently updated manually to ensure accurate alignment with the recorded audio and the screen recordings. The transcripts, audio, and screen recording files were stored confidentially on a secure server at Utrecht University. Upon completion of the investigation, the audio and screen recording files will be expeditiously deleted.

Certain quotes have been translated as precisely as possible from Dutch to English for inclusion in this thesis. Grammatical modifications have been made in a few quotes to ensure comprehensibility.

The NVivo analysis software tool was used to code the data [83]. Open coding [84] was employed to code the data. The classification process involved selecting and assigning transcribed passages to relevant categories.

In the following chapter, the outcomes of the evaluation sessions will be discussed.

8. Results Evaluation Sessions

This section presents the results of the evaluation sessions. First, we will outline the completion of the training by each participant (Paragraph 8.1). Subsequently, we shall discuss the outcomes obtained from the tasks they were asked to complete (Paragraph 8.2). Then, we shall assess the decision task's feasibility using the current tool (Paragraph 8.3). In addition, we will discuss the outcomes of the brief interview, highlighting both the strengths and opportunities for further improvements (Paragraph 8.4). Finally, an overview of possible future improvements will be provided (Paragraph 8.5).

8.1 Training Phase

All participants successfully completed the training without additional tasks required beyond the initial assessment. Two stopping rules were implemented during the training, whereby participants were given new tasks if they were unable to effectively complete the assigned task on the first attempt. However, this did not turn out to be necessary for any of the participants.

8.2 Judgement Tasks

All participants were able to complete the simpler tasks, which involved identifying the category of incidents, indicating where an incident will take place, how many people are expected to attend and so on.

However, it was discovered that not everyone understood the certainty score. For instance, participant P01 believed that the entire bar was part of the certainty score, leading to 60% of the model's answer being uncertain, 20% certain, and the remaining 20% very certain. However, the bar graph also displays a line indicating the category in which the certainty score falls, as there are three distinct levels of certainty. This was not noted by P01. On the other hand, participants P02 and P03 were able to correctly identify the model's certainty score.

Although participants could correctly identify the features utilised by the model for incident classification, they appear to lack an understanding of feature importance. When completing the tasks, participants seemed to focus more on the overall score and assumed uncertainty in the model due to a lack of information in the description.

8.3 Decision Task

The calendar view, which is available in both weekly and monthly formats, makes it easy to compare different incidents. All three participants commented that they really liked being able to see at a glance which incidents were happening at the same time. Furthermore, the risk classification visualisation expedites evaluation to determine where more effort is necessary on a given day.

The map proved highly beneficial to all three participants. It also facilitates the examination of whether two occurrences on the same day entail a mutual danger and necessitate additional police deployment.

For participant P01, in the design, the model would only be a convenient way to provide a quick indication:

> *"The model does provide support, but is certainly not decisive.""* - P01 Events/demonstration coordinator

An important distinction between the participants is that P01 and P03 serve as the primary decision makers in the realm of public order and safety, while P02 is responsible for gathering information for the decision makers. Conversely, P01 and P03 expressed the desire for greater reliance on human-provided information. Notably, during the evaluation session, P02 trusted the model considerably, especially when the level of certainty was high. This inclination may be attributed to the limited scope of the tool's examples.

It should be verified that, for example, if a model can find a lot of old information about an organisation and is therefore very confident that there is no risk, P01 and P03 will also see the model as leading, or they will still have a classification of people they would like.

8.4 Strengths, weaknesses and opportunities

In this section, we will evaluate both the strengths and weaknesses of the given system. Firstly, we identify the positive points in Paragraph 8.4.1. This is followed by a discussion of the possible areas for improvement in Paragraph 8.4.2.

8.4.1 Strengths

Firstly, the primary benefit of the design is its ability to offer an overview of upcoming incidents. The calendar format contributes to this clarity, making it effortless to compare days or incidents. All three participants confirmed this advantage.

The information can be read by clicking on it and it is easy to return to the full overview, making working with this version much more efficient compared to the current version of the police calendar. In current version, this is more difficult as users are directed to a new screen and have to switch between the information and the overview repeatedly.

Furthermore, all three participants found it very useful to be able to hover over an incident to quickly see some additional information. However, in the weekly calendar, this is only visible above an incident and it would be better to see it at the point of the mouse.

The colours make it easy to differentiate between different types of incidents.

8.4.2 Weaknesses

Not all participants found the colour scheme and stripe design to be convenient. P01, for instance, believed that the use of green may convey that everything is acceptable and there is no need for further action.

Furthermore, participants P02 and P03 expressed that the stripes elicited a contrasting interpretation in their perspective. Nonetheless, participant P01 and experienced no difficulty in interpreting the stripes in the current form.

The tool constantly displays a legend indicating the colour associated with incident types and the corresponding encoding for security scores. However, the legend does not provide information about the risk levels, which can only be accessed through a mouse movement. Additionally, the light grey colour, representing no classification given by humans, is not displayed. Participant P02 requests all information to be presented in the legend, including the levels which require a mouse movement to access.

However, each of the three sessions indicates that the currently available information on each incident is a significantly simplified version of reality. This is not just limited to the amount of information provided per incident but also the frequency of incidents occurring each day. Therefore, the advantages of implementing a model may not be apparent to the user. Further research should provide more realistic examples, after which another evaluation should be carried out to see if the findings remain the same.

8.5 Future improvements

Due to time constraints, another iteration of the design cannot be completed. However, potential points for future versions of the design will be provided in this section.

Firstly, the use of colours and stripes should be thoroughly reassessed. The sessions indicated that the colours can sometimes be confusing, as they can be interpreted in different ways.

Next, more realistic examples should be made so that they are more in line with reality. This needs to apply to both incident details and the frequency of incidents in a week/month. An option should also be given that can show the incidents per day or per year.

Furthermore, we have decided to focus this tool only on incidents in the month of October in the district of Stad-Utrecht, however, it can be extended to include all districts in the Netherlands and other dates. In this way, incidents on district borders could be discussed with neighbouring districts, which would lead to an optimisation of the deployment of resources.

Finally, we have not yet given the possibility to determine the deployment of units in this tool. The sessions highlighted the value decision makers placed on the ability to select the units deployed per incident within the tool. This facilitates comparison and possibly modification of deployment, particularly where an incident requires additional resources. The overview provides information on the location of each unit, enabling quick identification and adjustment in case of district issues.

Summary

The evaluation sessions indicate that the tool aids in decision making regarding public order and safety. However, the addition of the model seems to lack significant contribution in its current form, possibly due to the limited information available concerning the incidents within the current design.

In addition, there is a lack of consensus on the correct interpretation of the certainty score. This score does not seem to be taken into account by everyone when making decisions in the current design either.

The explanation of the model is partially comprehensible to users. They understand that certain features contribute to the classification. However, they do not seem to be able to recognise that some features contribute more to the classification than others. Future studies should investigate if more extensive information improves comprehension of the explanation and whether model utilisation offers added benefits in this scenario.

9. Discussion

In this Chapter, we provide a reflective analysis of our research. Furthermore, we will present the limitations of our research.

Our research aimed to effectively visualise local explanations of machine learning models for law enforcement decision makers. However, our search revealed that most visualization tools are designed to explain ML models to data scientists [25, 26, 27]. The distinction in their objectives when interacting with ML models necessitates different visualization approaches.

The utilization of ML models in law enforcement is imperative due to the need to analyze large, complex, and unstructured datasets to enforce the law [5]. Nevertheless, the foremost appropriate models in this field are the black-box models [45], which cannot explain their behaviour to users [45, 5, 32]. However, in high-stakes scenarios such as law enforcement, it is crucial that ML models are capable of explaining their decisions to the user. The models are used in high-stake environments and the decisions made with these models can have major consequences for both individuals and society [5, 23], especially when the models contain unnecessary bias, as illustrated by COMPAS [40].

9.1 Implications LIME and SHAP

Fortunately, post-hoc methods are available to explain the behaviour of black-box models [32]. A distinction is made between model-specific and model-agnostic methods [32]. This research focused on the two well-known model-agnostic post-hoc methods, Local Interpretable Model-Agnostic Explanations (LIME) [51] and SHapley Additive exPlanations (SHAP) [68]. The model-independent nature of these two XAI methods can provide developers and designers with a unified interface for the law enforcement domain, even if the ML model or underlying data changes [32].

Kauer et al.'s [70] research discovered that technical experts may struggle to interpret the standard LIME and SHAP visualisations. We therefore conducted research into how these two methods have been used in different domains and for different audiences. Our findings revealed that most tools were tailored to provide data scientists with insights on ML models [71, 72, 73, 74]. However, tools aimed at other users were not designed for the law enforcement domain. Nevertheless, related work has allowed us to identify several XAI interface design themes that explain local ML model decisions to users via LIME and SHAP.

9.2 Implications of Final Design

The findings from the literature review and the outcomes of the interviews with the public order and security decision makers led us to a number of key aspects for our design. After multiple revisions, our final design was implemented and subsequently evaluated.

The evaluation showed that the local explanation of the ML model could not be fully interpreted by the users. The features that contributed to the model's classification were recognised by the users, but they did not seem to recognise that some features contributed more than others. Nonetheless, this design proved more effective for decision makers than the current tool they employ, providing prospects for the future. Nevertheless, our design's oversimplified representation of reality may have impacted the study's outcomes.

9.3 Overall Limitations

First of all, the final design was evaluated by only three participants, which limits the generalisability of our conclusions.

Secondly, our sample solely consisted of decision makers within the domain of public order and safety in the district of Stad-Utrecht. As a result, we can state that our findings with regard to the needs and requirements that emerged from the research are only valid for the district of Stad-Utrecht. Further research is required to determine whether these requirements correspond with other districts across the Netherlands.

Furthermore, limited time meant that the incident data within the tool was insufficient to make informed decisions. While the design reflected actual incident information, it must be expanded to encompass more detail. A drawback of this is our inability to conduct a complete analysis of whether the explanations of the ML models are interpretable for the decision makers and how this information would ultimately influence the decision processes. Nonetheless, it is encouraging that the interface was predominantly perceived as intuitive and understandable for the relatively short time that users interacted with it.

Finally, in this research we have only focused on designing a visualisation that explains the local explanations of an ML model using LIME scores. Our literature review covered both LIME and SHAP. In addition, we only created one design for the explanation, so it could be that a different design is better for the interpretation by the users.

10. Conclusion & Prospects

In this study, we explored how to effectively visualise local explanations of ML models so that law enforcement decision makers can better understand them and integrate them into their decision making processes. The research was structured around the following research question:

RQ. How can local explanations of ML models be effectively visualised to enable law enforcement decision makers to better understand and incorporate them into their decision making process?

The main research question was investigated through three sub-questions. This Chapter concludes the research by answering each of the research questions in Section 10.1. Finally, Section 10.2 presents directions for future research.

10.1 Conclusion

In this section, we first present a summary of the main findings of each sub-question. Finally, we use the findings and inferences from these sub-questions to answer our main research question.

SQ1. What types of explanations are common in the field of XAI, especially relevant to law enforcement applications?

The common types of XAI explanations vary depending on the machine learning model used. While transparent models provide inherent explanations, black-box models do not. Nonetheless, black-box models are deemed appropriate for use in law enforcement due to their ability to analyse high-dimensional data sets. Therefore, it is necessary to explain the functioning of black-box models before they can be deployed. There are post-hoc techniques for this explanation, which can be divided into two categories: model-specific techniques that can only explain models from a specific group; and model-agnostic techniques that can explain models from all groups. Our research was centred on the latter because we had not determined beforehand which model we would concentrate on.

Subsequently, in Chapter 2, we established that visualisations are utilised to explain ML models. There are various techniques for creating such visualisations as identified in Chapter 3.

For clarity, we opted to concentrate our research on the two most commonly used model-agnostic post-hoc methods, LIME and SHAP. Chapter 3 subsequently examined existing tools that utilise these methods to explain ML models. It is worth noting that currently no tool has been developed with the specific purpose of explaining ML models in the context of law enforcement with visual aids. We noticed that numerous ML tools are in development for technical experts, but there are fewer options for non-technical experts who are domain specialists.

However, through this research, we identified several aspects for visualising the local explanation of an ML model, which we then used in our design:

1. Overview of Data: provide an overview of all items in the data set. Through a variety of interactions such as filtering, selecting, and ordering, the user is able to find specific and important items first.
2. Explanations on request: provide necessary additional information about what various features mean, or what the "normal" score would be for features using natural language.
3. Provide extra explanation in natural language: to ensure that the explanation does not depend solely on the user's interpretation, additional information, or labels, should be explained in natural language, if the user needs it.

SQ2. Who are the stakeholders that potentially interact with ML models in the law enforcement domain, and what are their needs related to the decision making process?

In Chapter 5, various stakeholders were identified who can interact with ML models in the law enforcement domain. We have focused on those involved in making decisions in the field of Public Order and Security. The key requirements for the decision making process were

- The source of the notification must be present.
- The identity of the organisation must be established in order to check whether there is any prior knowledge of it.
- The information collected should be shareable to enable it to be distributed effectively.
- There needs to be a means of distinguishing between different types of incident. This is because the specific combination of factors involved, such as the time of day and the target group, will require the police to vary their level of deployment.
- To achieve a full understanding of what is about to happen, the key aspects of the incident need to be presented in a way which enables all concerned to see what is going to happen in the coming week, month or year.

SQ3. How can the insights from SQ1 and SQ2 be utilised to develop an effective and interpretable visualisation of local explanations for decision makers in the law enforcement domain?

The insights from both the interviews and the literature review have resulted in the final list of requirements for the design:

- The source of the notification must be present.
- The identity of the organisation must be established in order to check whether there is any prior knowledge of it.
- The information collected should be shareable to enable it to be distributed effectively.
- There needs to be a means of distinguishing between different types of incident. This is because the specific combination of factors involved, such as the time of day and the target group, will require the police to vary their level of deployment.
- To achieve a full understanding of what is about to happen, the key aspects of the incident need to be presented in a way which enables all concerned to see what is going to happen in the coming week, month or year.
- To provide extra explanations for the local explanations of ML models in natural language.
- To provide explanations where necessary. In other words, by making all the necessary additional information visible at some point.
- To provide an overview of the full dataset.

We eventually translated this list into a design and after some iterations the final design emerged as described in chapter 6.

The evaluation study findings demonstrate that decision makers can locate the appropriate information they need to guide their decision making processes. However, it transpired that not everyone could interpret the certainty scores. Moreover, the evaluation also revealed that users were only partially able to understand the model's explanation. Users understood that certain features contribute to the model's risk classification, but they did not seem to realise that some features have a greater contribution to the classification than others. Additionally, the examples used in the tool were found to be overly simplistic, which could potentially impact the obtained outcomes.

RQ. How can local explanations of ML models be effectively visualised to enable law enforcement decision makers to better understand and incorporate them into their decision making process?

Our design for explaining the local behaviour has been demonstrated to benefit policy makers within the law enforcement sector, particularly those involved in public order and safety decision making. Nonetheless, we have observed limitations in users' ability to interpret the local explanations of the model accurately. The participants identified the features that contributed to the model's classification, but did not acknowledge the varying degrees of contribution.

Some users did not perceive the model as an enhancement to the existing design, possibly due to the design's limited examples of incidents. In reality, incident information is much more detailed and therefore, a more extensive illustration may be required for the model to be recognised as an enrichment.

Further study is required to determine if increasing the amount of examples can enhance the comprehension of local interpretations of machine learning models. Additionally, a method for clarifying the varying contributions of features is required to better explain the different contributions of features.

10.2 Future work

A number of potential opportunities for future work have emerged from this thesis. This section will discuss these directions for further research.

The first of these opportunities could lie in refining and improving the final design, which needs to become more specific. Our design was found to be a limited representation of reality. Initially, this was due to the relatively low number of incidents displayed per day. In reality, many more different incidents could occur simultaneously in a day, which would increase the complexity of the visualisation. Furthermore, the data offered within each individual incident was insufficient. In reality, there is a lot more information available to ultimately determine the decision to respond to an incident. Furthermore, in this design we only focused on incidents in the city of Utrecht for the month of October. However, including all districts in the Netherlands and gathering data on all potential incidents would further enhance the tool's effectiveness in evaluating deployment strategies. Finally, the tool's design could be improved by determining the deployment and visualising it in the tool.

Secondly, the user research conducted in this thesis can be used as a foundation to create visualisations to explain the local explanations of an ML model to decision makers in other law enforcement contexts.

Furthermore, this study ultimately focused only on the visualisation of LIME scores. Further research could explore the visualisation of Shapley (SHAP) scores. Alternatively, research could be conducted to determine which of the two methods is the most comprehensible to decision makers.

In addition, this research focused on the local explanation of an ML model. However, the literature review highlights the existence of global explanations. Further studies could explore how to visualise these explanations in an effective way for law enforcement decision makers.

Finally, future research could also look at other ways of visualising the local explanations and possibly compare them with this visualisation to investigate what the desired way of visualising these explanations is.

Bibliography

[1] Politiewet, "Wetten.nl - regeling - politiewet 2012 - BWBR0031788," 2023. Accessed at 30-07-2023, on: https://wetten.overheid.nl/BWBR0031788/2013-05-01.

[2] K. M. Hess, C. H. Orthmann, and H. L. Cho, *Introduction to law enforcement and criminal justice*. Cengage Learning, 2014.

[3] R. Meijer, D. Moolenaar, R. Choenni, S. Van den Braak, W. de Jongste, M. Akkermans, E. Moons, W. Schirm, R. Kessels, C. Netten, *et al.*, "Criminaliteit en rechtshandhaving 2021," *WODC*, 2022.

[4] O. Solakoglu, "Trust in Police: A Comparative Study of Belgium and The Netherlands.," *International Journal of Criminal Justice Sciences*, vol. 11, no. 1, 2016.

[5] F. Dechesne, V. Dignum, L. Zardiashvili, and J. Bieger, "AI & ethics at the police: Towards responsible use of artificial intelligence in the Dutch Police," *AI & Ethics at the Police: Towards Responsible use of Artificial Intelligence in the Dutch Police*, 2019.

[6] Europol, "Netherlands | Europol," 10 2020. Accessed at 01-08-2023, on: https://www.europol.europa.eu/partners-collaboration/member-states/netherlands.

[7] OHCHR, "Code of Conduct for Law Enforcement Officials," 12 1979. https://www.ohchr.org/en/instruments-mechanisms/instruments/code-conduct-law-enforcement-officials.

[8] Interpol, "How INTERPOL supports the Netherlands to tackle international crime," 2023. Accessed at 01-08-2023, on: https://www.interpol.int/Who-we-are/Member-countries/Europe/NETHERLANDS.

[9] K. Glasgow, "Chapter 4 - Big Data and Law Enforcement: Advances, Implications, and Lessons from an Active Shooter Case Study," in *Application of Big Data for National Security*, pp. 39–54, Butterworth-Heinemann, 2015. https://doi.org/10.1016/B978-0-12-801967-2.00004-5.

[10] S. L. Garfinkel, "Digital forensics research: The next 10 years," *Digital Investigation*, vol. 7, pp. S64–S73, 2010. https://doi.org/10.1016/j.diin.2010.05.009.

[11] S. Brayne, "The Criminal Law and Law Enforcement Implications of Big Data," *Annual Review of Law and Social Science*, vol. 14, no. 1, pp. 293–308, 2018. doi: 10.1146/annurev-lawsocsci.

[12] W. Hardyns and A. Rummens, "Predictive Policing as a New Tool for Law Enforcement? Recent Developments and Challenges," *European Journal on Criminal Policy and Research*, vol. 24, pp. 201–218, 2018. https://doi.org/10.1007/s10610-017-9361-2.

[13] G. Kejela, R. M. Esteves, and C. Rong, "Predictive analytics of sensor data using distributed machine learning techniques," in *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, pp. 626–631, 2014. doi: 10.1109/CloudCom.2014.44.

[14] I. H. Sarker, "Machine learning: algorithms, Real-World applications and research directions," *SN Computer Science*, vol. 2, 3 2021. https://doi.org/10.1007/s42979-021-00592-x.

[15] L. B. Moses and J. Chan, "Using big data for legal and law enforcement decisions: Testing the new tools," *University of New South Wales Law Journal, the*, vol. 37, no. 2, pp. 643–678, 2014. https://search.informit.org/doi/10.3316/informit.613165001799453.

[16] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren, "The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations," *Computer Graphics Forum*, vol. 39, no. 3, pp. 713–756, 2020. https://doi.org/10.1111/cgf.14034.

[17] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, 2017. https://doi.org/10.1016/j.neucom.2017.01.026.

[18] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, (New York, NY, USA), p. 1–18, Association for Computing Machinery, 2018. https://doi.org/10.1145/3173574.3174156.

[19] S. Mohseni, N. Zarei, and E. D. Ragan, "A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems," *ACM Transactions on Interactive Intelligent Systems*, vol. 11, no. 3-4, 2021. https://doi.org/10.1145/3387166.

[20] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017. https://doi.org/10.1089/big.2016.0047.

[21] A. Preece, D. Harborne, D. Braines, R. Tomsett, and S. Chakraborty, "Stakeholders in Explainable AI," *arXiv preprint arXiv:1810.00184*, 2018. http://arxiv.org/abs/1810.00184.

[22] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady, "explAIner: A visual analytics framework for interactive and explainable machine learning," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 1064–1074, 2019. doi: 10.1109/TVCG.2019.2934629.

[23] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature machine intelligence*, vol. 1, no. 5, pp. 206–215, 2019. https://doi.org/10.1038/s42256-019-0048-x.

[24] E. Bertini and D. Lalanne, "Surveying the Complementary Role of Automatic Data Analysis and Visualization in Knowledge Discovery," in *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration*, (New York, NY, USA), p. 12–20, Association for Computing Machinery, 2009. https://doi.org/10.1145/1562849.1562851.

[25] F. Sperrle, M. El-Assady, G. Guo, D. H. Chau, A. Endert, and D. Keim, "Should We Trust (X)AI? Design Dimensions for Structured Experimental Evaluations," *arXiv preprint arXiv:2009.06433*, 2020.

[26] E. Dimara, H. Zhang, M. Tory, and S. Franconeri, "The unmet data visualization needs of decision makers within organizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 12, pp. 4101–4112, 2022. doi: 10.1109/TVCG.2021.3074023.

[27] T. Miller, P. Howe, and L. Sonenberg, "Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences," *arXiv preprint arXiv:1712.00547*, 2017.

[28] E. Dimara and J. Stasko, "A critical reflection on visualization research: Where do decision making tasks hide?," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 1128–1138, 2022. doi: 10.1109/TVCG.2021.3114813.

[29] S. Liu, X. Wang, M. Liu, and J. Zhu, "Towards better analysis of machine learning models: A visual analytics perspective," *Visual Informatics*, vol. 1, no. 1, pp. 48–56, 2017. https://doi.org/10.1016/j.visinf.2017.01.006.

[30] F. Sperrle, M. El-Assady, G. Guo, R. Borgo, D. H. Chau, A. Endert, and D. Keim, "A Survey of Human-Centered Evaluations in Human-Centered Machine Learning," *Computer Graphics Forum*, vol. 40, no. 3, pp. 543–568, 2021. https://doi.org/10.1111/cgf.14329.

[31] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty, "Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems," 2018.

[32] C. Molnar, *Interpretable machine learning*. Lulu. com, 8 2023. Accessed at 10-07-2023, on https://christophm.github.io/interpretable-ml-book/.

[33] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018. doi: 10.1109/ACCESS.2018.2870052.

[34] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," *arXiv preprint arXiv:1702.08608*, 2017.

[35] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13-17, 2019*, pp. 1078–1088, International Foundation for Autonomous Agents and Multiagent Systems, 2019.

[36] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, "Evaluating XAI: A comparison of rule-based and example-based explanations," *Artificial Intelligence*, vol. 291, p. 103404, 2021. doi: 10.1016/j.artint.2020.103404.

[37] M. Sedlmair, M. Meyer, and T. Munzner, "Design study methodology: Reflections from the trenches and the stacks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2431–2440, 2012. doi: 10.1109/TVCG.2012.213.

[38] P. Nederland, "Organisatiestructuur Nederlandse politie." Accessed at 10-09-2023, https://www.politie.nl/informatie/organisatiestructuur-nederlandse-politie.html.

[39] A. F. Osman, "Radiation oncology in the era of big data and machine learning for precision medicine," *Artificial Intelligence-Applications in Medicine and Biology. IntechOpen*, pp. 41–70, 2019.

[40] C. Rudin and B. Ustun, "Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice," *Interfaces*, vol. 48, no. 5, pp. 449–466, 2018. https://doi.org/10.1287/inte.2018.0957.

[41] T. Brennan, W. Dieterich, and B. Ehret, "Evaluating the predictive validity of the COMPAS risk and needs assessment system," *Criminal Justice and behavior*, vol. 36, no. 1, pp. 21–40, 2009. https://doi.org/10.1177/0093854808326545.

[42] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, and R. Krishna, "Explanations Can Reduce Overreliance on AI Systems During Decision-Making," *Proc. ACM Hum.-Comput. Interact.*, vol. 7, apr 2023. https://doi.org/10.1145/3579605.

[43] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, apr 2021. https://doi.org/10.1145/3449287.

[44] D. Gunning, "Explainable artificial intelligence (XAI)," 2023. Accessed at 20-07-2023, on https://www.darpa.mil/program/explainable-artificial-intelligence.

[45] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020. https://doi.org/10.1016/j.inffus.2019.12.012.

[46] Z. C. Lipton, “The Mythos of Model Interpretability,” *Commun. ACM*, vol. 61, p. 36–43, sep 2018. doi: 10.1145/3233231.

[47] A. A. Freitas, “Comprehensible Classification Models: A Position Paper,” *SIGKDD Explor. Newsl.*, vol. 15, p. 1–10, mar 2014. https://doi.org/10.1145/2594473.2594475.

[48] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens, “An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models,” *Decision Support Systems*, vol. 51, no. 1, pp. 141–154, 2011. https://doi.org/10.1016/j.dss.2010.12.003.

[49] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A Survey of Methods for Explaining Black Box Models,” *ACM Comput. Surv.*, vol. 51, aug 2018. https://doi.org/10.1145/3236009.

[50] Z. Chen, F. Xiao, F. Guo, and J. Yan, “Interpretable machine learning for building energy management: A state-of-the-art review,” *Advances in Applied Energy*, vol. 9, p. 100123, 2023. https://doi.org/10.1016/j.adapen.2023.100123.

[51] M. T. Ribeiro, S. Singh, and C. Guestrin, “"Why Should I Trust You?": Explaining the Predictions of Any Classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 1135–1144, 2016. https://doi.org/10.1145/2939672.2939778.

[52] A. Kirsch, “Explain to whom? Putting the user in the center of explainable AI,” in *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2017)*, 2017.

[53] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. van Moorsel, “The Relationship between Trust in AI and Trustworthy Machine Learning Technologies,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* ’20, (New York, NY, USA), p. 272–283, Association for Computing Machinery, 2020. https://doi.org/10.1145/3351095.3372834.

[54] J. Vermeulen, G. Vanderhulst, K. Luyten, and K. Coninx, “PervasiveCrystal: Asking and Answering Why and Why Not Questions about Pervasive Computing Applications,” in *2010 Sixth International Conference on Intelligent Environments*, pp. 271–276, 2010. 10.1109/IE.2010.56.

[55] C. J. Cai, J. Jongejan, and J. Holbrook, “The Effects of Example-Based Explanations in a Machine Learning Interface,” in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI ’19, (New York, NY, USA), p. 258–262, Association for Computing Machinery, 2019. https://doi.org/10.1145/3301275.3302289.

[56] B. Y. Lim, A. K. Dey, and D. Avrahami, “Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’09, (New York, NY, USA), p. 2119–2128, Association for Computing Machinery, 2009. https://doi.org/10.1145/1518701.1519023.

[57] R. Kocielnik, S. Amershi, and P. N. Bennett, "Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, (New York, NY, USA), p. 1–14, Association for Computing Machinery, 2019. https://doi.org/10.1145/3290605.3300641.

[58] B. Y. Lim, Q. Yang, A. M. Abdul, and D. Wang, "Why these explanations? Selecting intelligibility types for explanation goals.," in *IUI Workshops*, 2019.

[59] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, and M. Terry, "Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, (New York, NY, USA), p. 1–14, Association for Computing Machinery, 2019. https://doi.org/10.1145/3290605.3300234.

[60] F. J. Anscombe, "Graphs in statistical analysis," *The American Statistician*, vol. 27, no. 1, pp. 17–21, 1973. doi: 10.1080/00031305.1973.10478966.

[61] S. K. Card, J. D. Mackinlay, and B. Shneiderman, "Using vision to think," *Readings in information visualization: using vision to think*, pp. 579–581, 1999.

[62] T. Munzner, *Visualization analysis and design*. CRC press, 2014.

[63] B. La Rosa, G. Blasilli, R. Bourqui, D. Auber, G. Santucci, R. Capobianco, E. Bertini, R. Giot, and M. Angelini, "State of the Art of Visual Analytics for eXplainable Deep Learning," *Computer Graphics Forum*, vol. 42, no. 1, pp. 319–355, 2023. https://doi.org/10.1111/cgf.14733.

[64] J. J. Van Wijk, "The value of visualization," in *VIS 05. IEEE Visualization, 2005.*, pp. 79–86, 2005. doi: 10.1109/VISUAL.2005.1532781.

[65] K. Kucher and A. Kerren, "Text visualization techniques: Taxonomy, visual survey, and community insights," in *2015 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 117–121, 2015. doi: 10.1109/PACIFICVIS.2015.7156366.

[66] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," *Journal of the American statistical association*, vol. 79, no. 387, pp. 531–554, 1984. doi: 10.1080/01621459.1984.10478080.

[67] E. Dimara and C. Perin, "What is Interaction for Data Visualization?," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, pp. 119–129, 1 2020. doi: 10.1109/TVCG.2019.2934283.

[68] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[69] M. Chromik, "Making SHAP Rap: Bridging local and global insights through interaction and narratives," in *Human-Computer Interaction–INTERACT*

2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18, pp. 641–651, Springer, 2021. https://doi.org/10.1179/000870403235002042.

[70] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, (New York, NY, USA), p. 1–14, Association for Computing Machinery, 2020. https://doi.org/10.1145/3313831.3376219.

[71] L. Meng, S. van den Elzen, and A. Vilanova, "ModelWise: Interactive Model Comparison for Model Diagnosis, Improvement and Selection," *Computer Graphics Forum*, vol. 41, no. 3, pp. 97–108, 2022. https://doi.org/10.1111/cgf.14525.

[72] Z. Li, X. Wang, W. Yang, J. Wu, Z. Zhang, Z. Liu, M. Sun, H. Zhang, and S. Liu, "A unified understanding of deep nlp models for text classification," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 12, pp. 4980–4994, 2022. doi: 10.1109/TVCG.2022.3184186.

[73] M. Angelini, G. Blasilli, S. Lenti, and G. Santucci, "A Visual Analytics Conceptual Framework for Explorable and Steerable Partial Dependence Analysis," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–16, 2023. doi: 10.1109/TVCG.2023.3263739.

[74] X. Zhao, W. Cui, Y. Wu, H. Zhang, H. Qu, and D. Zhang, "Oui! Outlier Interpretation on Multi-dimensional Data via Visual Analytics," *Computer Graphics Forum*, vol. 38, no. 3, pp. 213–224, 2019. https://doi.org/10.1111/cgf.13683.

[75] L. E. Resck, J. R. Ponciano, L. G. Nonato, and J. Poco, "Legalvis: Exploring and inferring precedent citations in legal documents," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 6, pp. 3105–3120, 2023. doi: 10.1109/TVCG.2022.3152450.

[76] F. Cheng, D. Liu, F. Du, Y. Lin, A. Zytek, H. Li, H. Qu, and K. Veeramachaneni, "Vbridge: Connecting the dots between features and data to explain healthcare models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 378–388, 2022. doi: 10.1109/TVCG.2021.3114836.

[77] L. Chen, Y. Ouyang, H. Zhang, S. Hong, and Q. Li, "RISeer: Inspecting the Status and Dynamics of Regional Industrial Structure via Visual Analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 1070–1080, 2023. doi: 10.1109/TVCG.2022.3209351.

[78] C. Zhang, X. Wang, C. Zhao, Y. Ren, T. Zhang, Z. Peng, X. Fan, X. Ma, and Q. Li, "PromotionLens: Inspecting Promotion Strategies of Online E-commerce via Visual Analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 1, pp. 767–777, 2023. doi: 10.1109/TVCG.2022.3209440.

[79] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proceedings 1996 IEEE symposium on visual languages*, pp. 336–343, 1996. doi: 10.1109/VL.1996.545307.

[80] B. B. Frey, *The SAGE encyclopedia of educational research, measurement, and evaluation*. Sage Publications, 2018.

[81] I. Etikan, S. A. Musa, R. S. Alkassim, *et al.*, “Comparison of convenience sampling and purposive sampling,” *American journal of theoretical and applied statistics*, vol. 5, no. 1, pp. 1–4, 2016.

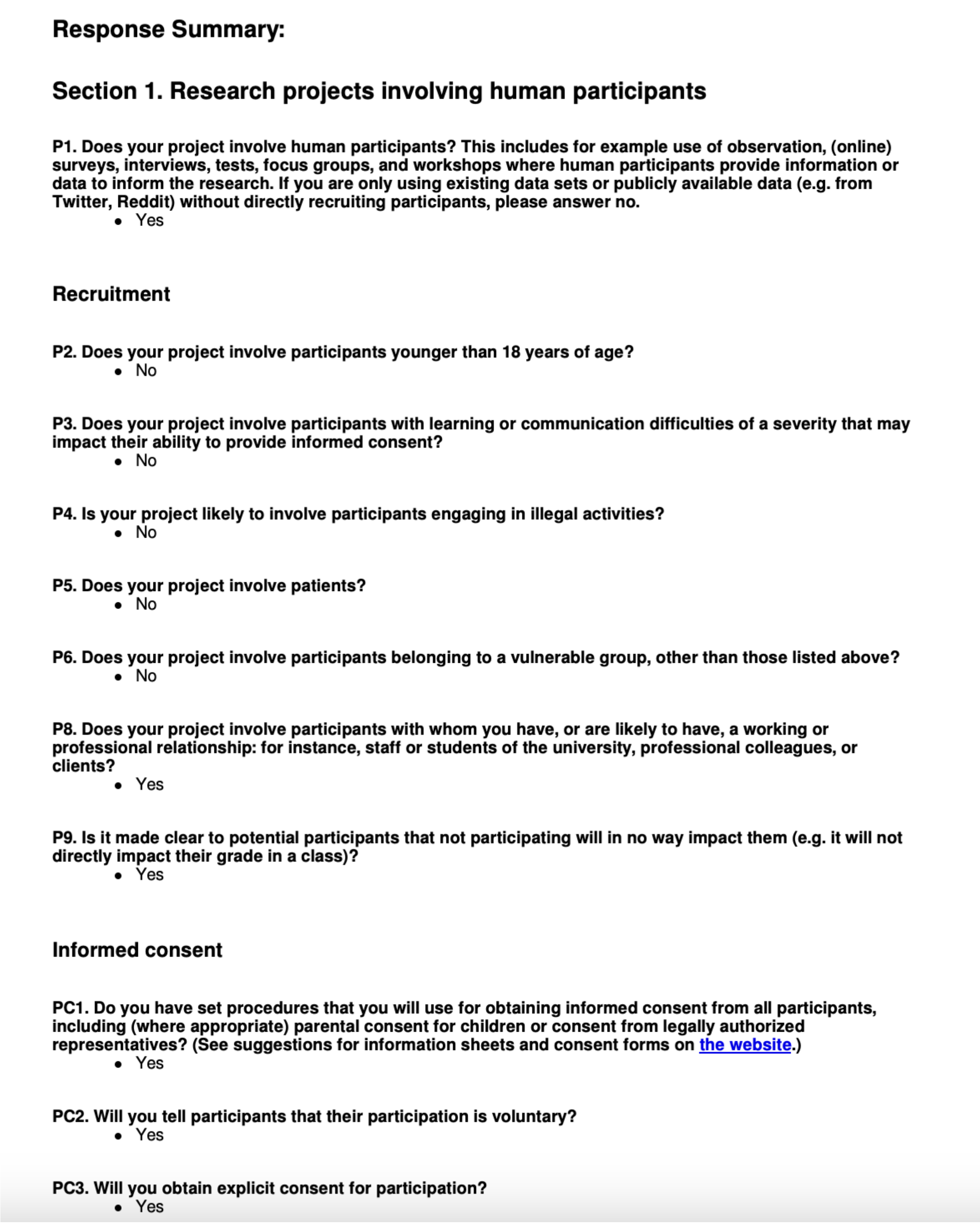
[82] A. Bryman, *Social research methods*. Oxford university press, 2016.

[83] Lumivero, “NVivo - Lumivero,” 2023. https://lumivero.com/products/nvivo/.

[84] H. Boeije and I. Bleijenbergh, “Analyseren in kwalitatief onderzoek (3de editie),” *Boom Lemma*, 2019.

[85] M. Harrower and C. A. Brewer, “ColorBrewer. org: an online tool for selecting colour schemes for maps,” *The Cartographic Journal*, vol. 40, no. 1, pp. 27–37, 2003. https://doi.org/10.1179/000870403235002042.

A. Quick Scan report



Response Summary:

Section 1. Research projects involving human participants

P1. Does your project involve human participants? This includes for example use of observation, (online) surveys, interviews, tests, focus groups, and workshops where human participants provide information or data to inform the research. If you are only using existing data sets or publicly available data (e.g. from Twitter, Reddit) without directly recruiting participants, please answer no.

- Yes

Recruitment

P2. Does your project involve participants younger than 18 years of age?

- No

P3. Does your project involve participants with learning or communication difficulties of a severity that may impact their ability to provide informed consent?

- No

P4. Is your project likely to involve participants engaging in illegal activities?

- No

P5. Does your project involve patients?

- No

P6. Does your project involve participants belonging to a vulnerable group, other than those listed above?

- No

P8. Does your project involve participants with whom you have, or are likely to have, a working or professional relationship: for instance, staff or students of the university, professional colleagues, or clients?

- Yes

P9. Is it made clear to potential participants that not participating will in no way impact them (e.g. it will not directly impact their grade in a class)?

- Yes

Informed consent

PC1. Do you have set procedures that you will use for obtaining informed consent from all participants, including (where appropriate) parental consent for children or consent from legally authorized representatives? (See suggestions for information sheets and consent forms on the website.)

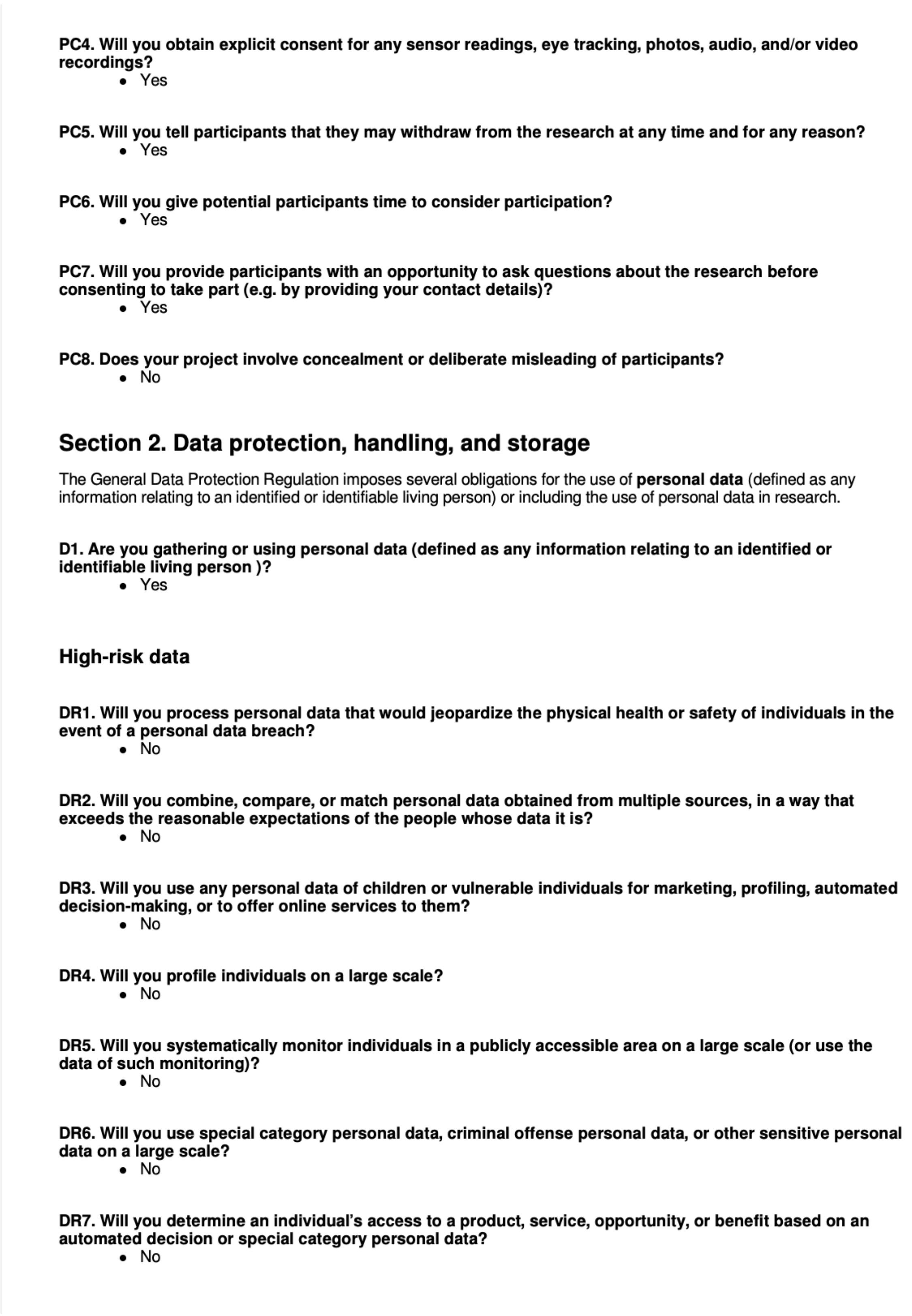
- Yes

PC2. Will you tell participants that their participation is voluntary?

- Yes

PC3. Will you obtain explicit consent for participation?

- Yes



PC4. Will you obtain explicit consent for any sensor readings, eye tracking, photos, audio, and/or video recordings?

- Yes

PC5. Will you tell participants that they may withdraw from the research at any time and for any reason?

- Yes

PC6. Will you give potential participants time to consider participation?

- Yes

PC7. Will you provide participants with an opportunity to ask questions about the research before consenting to take part (e.g. by providing your contact details)?

- Yes

PC8. Does your project involve concealment or deliberate misleading of participants?

- No

Section 2. Data protection, handling, and storage

The General Data Protection Regulation imposes several obligations for the use of **personal data** (defined as any information relating to an identified or identifiable living person) or including the use of personal data in research.

D1. Are you gathering or using personal data (defined as any information relating to an identified or identifiable living person)?

- Yes

High-risk data

DR1. Will you process personal data that would jeopardize the physical health or safety of individuals in the event of a personal data breach?

- No

DR2. Will you combine, compare, or match personal data obtained from multiple sources, in a way that exceeds the reasonable expectations of the people whose data it is?

- No

DR3. Will you use any personal data of children or vulnerable individuals for marketing, profiling, automated decision-making, or to offer online services to them?

- No

DR4. Will you profile individuals on a large scale?

- No

DR5. Will you systematically monitor individuals in a publicly accessible area on a large scale (or use the data of such monitoring)?

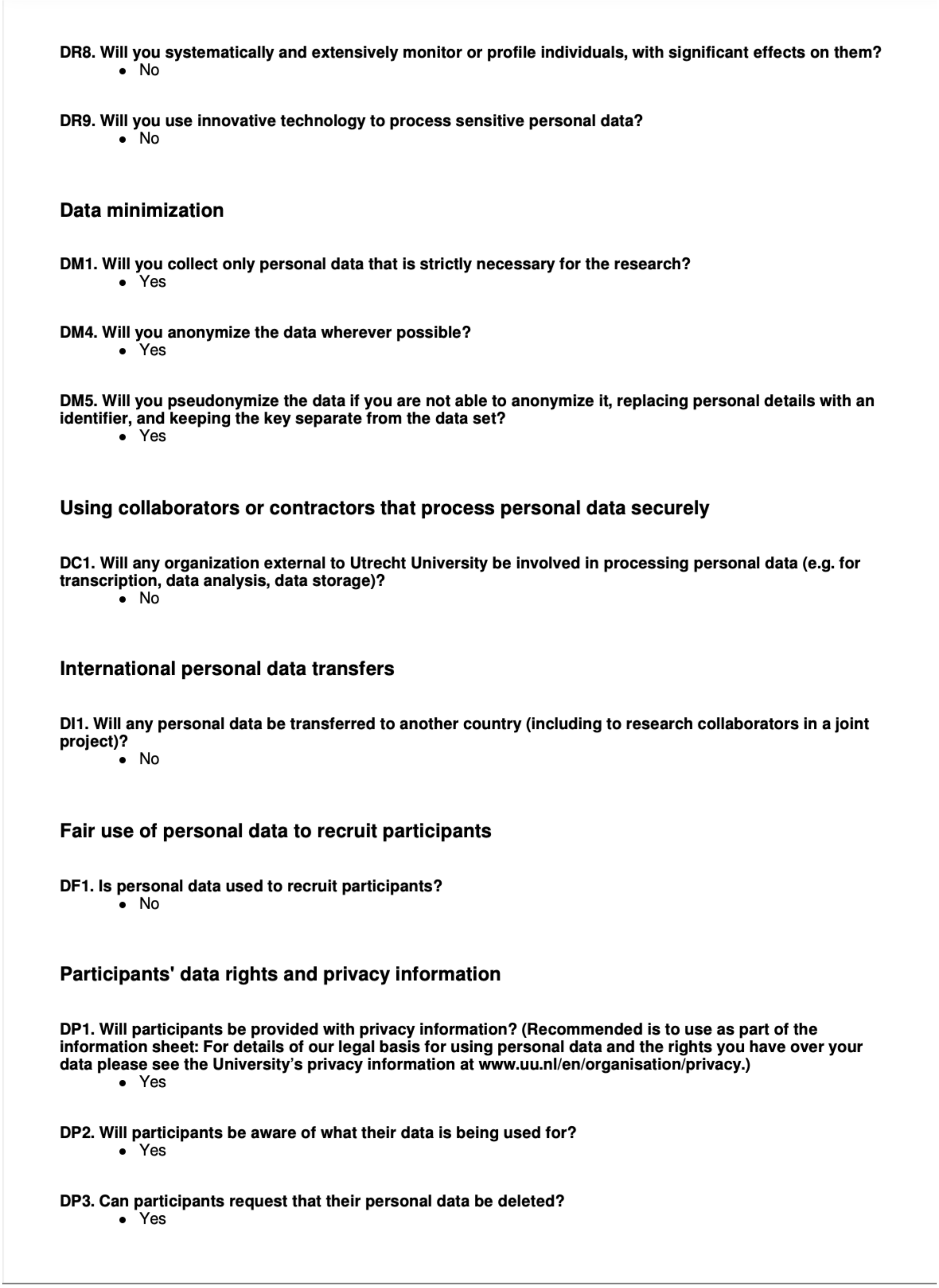
- No

DR6. Will you use special category personal data, criminal offense personal data, or other sensitive personal data on a large scale?

- No

DR7. Will you determine an individual’s access to a product, service, opportunity, or benefit based on an automated decision or special category personal data?

- No



DR8. Will you systematically and extensively monitor or profile individuals, with significant effects on them?

- No

DR9. Will you use innovative technology to process sensitive personal data?

- No

Data minimization

DM1. Will you collect only personal data that is strictly necessary for the research?

- Yes

DM4. Will you anonymize the data wherever possible?

- Yes

DM5. Will you pseudonymize the data if you are not able to anonymize it, replacing personal details with an identifier, and keeping the key separate from the data set?

- Yes

Using collaborators or contractors that process personal data securely

DC1. Will any organization external to Utrecht University be involved in processing personal data (e.g. for transcription, data analysis, data storage)?

- No

International personal data transfers

DI1. Will any personal data be transferred to another country (including to research collaborators in a joint project)?

- No

Fair use of personal data to recruit participants

DF1. Is personal data used to recruit participants?

- No

Participants' data rights and privacy information

DP1. Will participants be provided with privacy information? (Recommended is to use as part of the information sheet: For details of our legal basis for using personal data and the rights you have over your data please see the University's privacy information at www.uu.nl/en/organisation/privacy.)

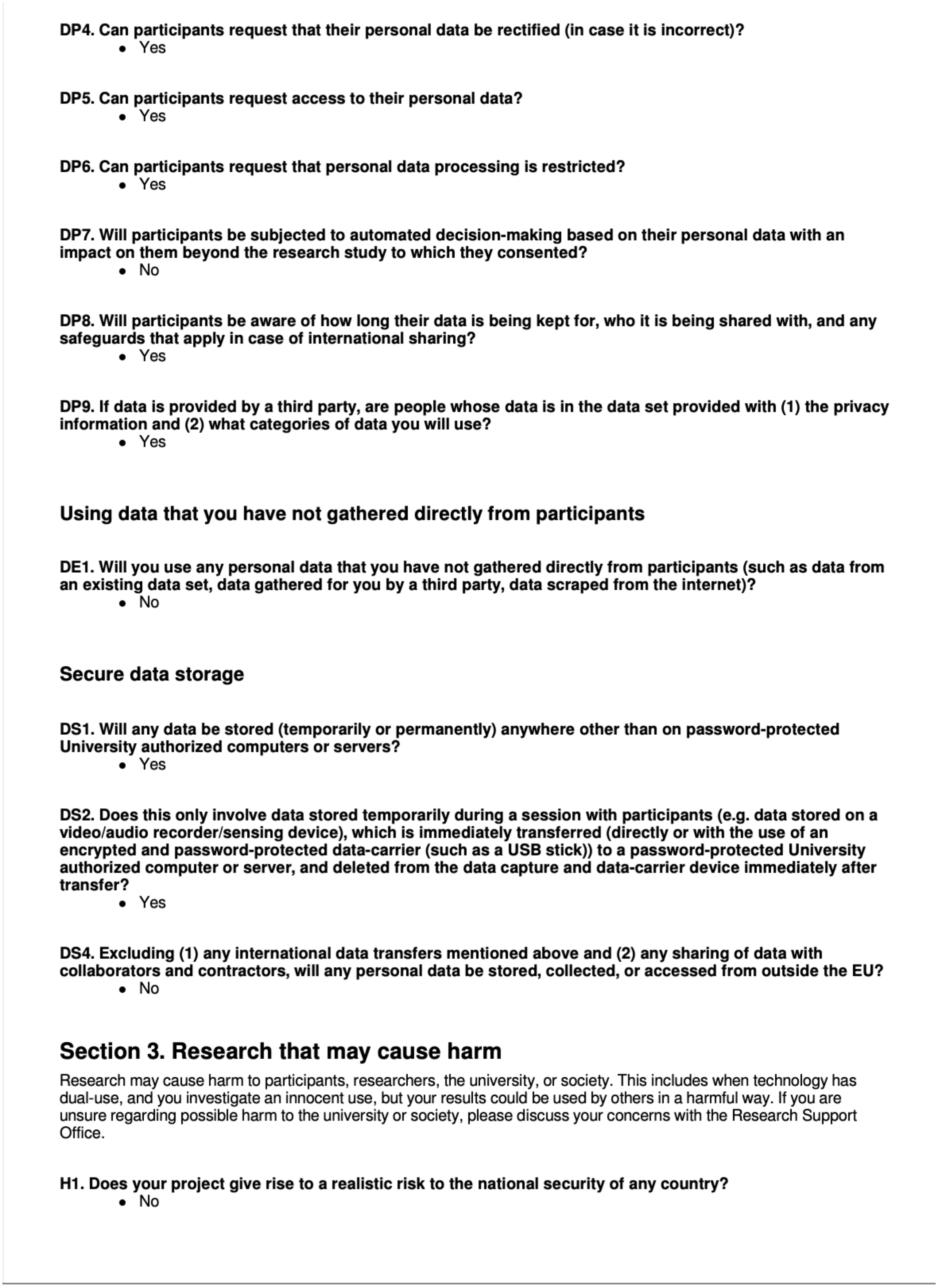
- Yes

DP2. Will participants be aware of what their data is being used for?

- Yes

DP3. Can participants request that their personal data be deleted?

- Yes



DP4. Can participants request that their personal data be rectified (in case it is incorrect)?

- Yes

DP5. Can participants request access to their personal data?

- Yes

DP6. Can participants request that personal data processing is restricted?

- Yes

DP7. Will participants be subjected to automated decision-making based on their personal data with an impact on them beyond the research study to which they consented?

- No

DP8. Will participants be aware of how long their data is being kept for, who it is being shared with, and any safeguards that apply in case of international sharing?

- Yes

DP9. If data is provided by a third party, are people whose data is in the data set provided with (1) the privacy information and (2) what categories of data you will use?

- Yes

Using data that you have not gathered directly from participants

DE1. Will you use any personal data that you have not gathered directly from participants (such as data from an existing data set, data gathered for you by a third party, data scraped from the internet)?

- No

Secure data storage

DS1. Will any data be stored (temporarily or permanently) anywhere other than on password-protected University authorized computers or servers?

- Yes

DS2. Does this only involve data stored temporarily during a session with participants (e.g. data stored on a video/audio recorder/sensing device), which is immediately transferred (directly or with the use of an encrypted and password-protected data-carrier (such as a USB stick)) to a password-protected University authorized computer or server, and deleted from the data capture and data-carrier device immediately after transfer?

- Yes

DS4. Excluding (1) any international data transfers mentioned above and (2) any sharing of data with collaborators and contractors, will any personal data be stored, collected, or accessed from outside the EU?

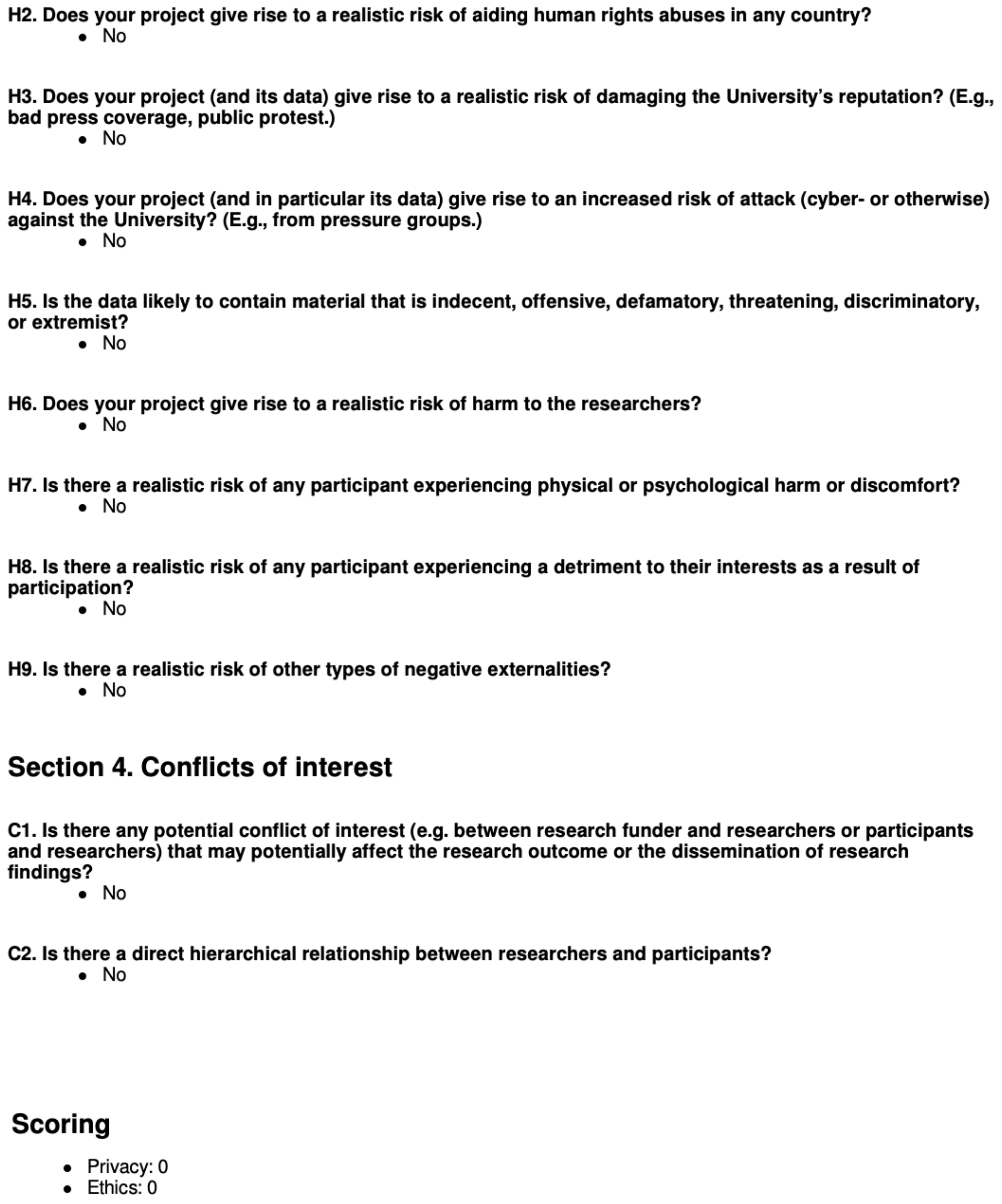
- No

Section 3. Research that may cause harm

Research may cause harm to participants, researchers, the university, or society. This includes when technology has dual-use, and you investigate an innocent use, but your results could be used by others in a harmful way. If you are unsure regarding possible harm to the university or society, please discuss your concerns with the Research Support Office.

H1. Does your project give rise to a realistic risk to the national security of any country?

- No



H2. Does your project give rise to a realistic risk of aiding human rights abuses in any country?

- No

H3. Does your project (and its data) give rise to a realistic risk of damaging the University's reputation? (E.g., bad press coverage, public protest.)

- No

H4. Does your project (and in particular its data) give rise to an increased risk of attack (cyber- or otherwise) against the University? (E.g., from pressure groups.)

- No

H5. Is the data likely to contain material that is indecent, offensive, defamatory, threatening, discriminatory, or extremist?

- No

H6. Does your project give rise to a realistic risk of harm to the researchers?

- No

H7. Is there a realistic risk of any participant experiencing physical or psychological harm or discomfort?

- No

H8. Is there a realistic risk of any participant experiencing a detriment to their interests as a result of participation?

- No

H9. Is there a realistic risk of other types of negative externalities?

- No

Section 4. Conflicts of interest

C1. Is there any potential conflict of interest (e.g. between research funder and researchers or participants and researchers) that may potentially affect the research outcome or the dissemination of research findings?

- No

C2. Is there a direct hierarchical relationship between researchers and participants?

- No

Scoring

- Privacy: 0
- Ethics: 0

Figure A.1: *Anonomysised Quick Scan*

B. Consent form

Title of research study: UX/UI design Explabox

Dear participant,

You will be asked to participate in a scientific study. Before you decide to participate, it is important that you know what participating in this study entails. Please read the information below carefully.

Principal Investigators: Marcel Robeer, marcel.robeer@politie.nl and Evanthia Dimara, e.dimara@uu.nl.

Student researcher: Kim van Genderen, m.s.vangenderen@students.uu.nl.

Supported by: This research is supported by the National Police Lab AI, nationaalpolitielab@uu.nl.

Important information about this study:
To help you decide whether you want to participate in the study, a brief summary follows. More detailed information is given further down the form.

- The purpose of the research is to find out what the difficulties are in your daily decision-making and information transfer, so that I can understand how better solutions can be designed to help you.
- We expect this research study to last about 1 hour.
- Your participation in this interview does not involve greater risks than you would encounter in everyday life. There are no expected risks beyond the risks that may be associated with computing.
- Based on a Quick Scan Ethics and Privacy, this research has received approval from the Research Institute of Information and Computing Sciences. If you have a complaint or comment about the way this investigation was conducted, you can send an email to ics-ethics@uu.nl. For questions or complaints about the processing of personal data, you can send an e-mailto the Data Protection Officer of Utrecht University via privacy-beta@uu.nl.

Why am I being asked to participate?
We ask you to participate in this investigation because you are between 18 and 65 years old, and are involved in decision-making as part of your position within the police.

What do I need to know about participating in this study?

- Participation in this study is entirely voluntary, whether or not you participate is up to you.
- You can choose not to participate.
- You can agree to participate and change your mind later.
- Your decision will not be held against you.
- You can ask all the questions you want before you decide.
- You don't have to answer questions you don't want to answer.
- You can stop at any time during the study. This can be done after collecting the data, but also during the collection of the data.

What happens if I say, "Yes, I want to participate in this study"?

- The interview is recorded (audio). You agree to this if you agree to participate in the study. The recordings are only used for research purposes.
- During the first part of the interview, you will be asked about your previous experiences and current working practices.
- Finally, you will be asked to complete a short demographic survey with questions such as your years of experience in your position within the police force.

Is there any way that participating in this study could be bad for me?

Your participation in this interview does not involve greater risks than you would encounter in everyday life. There are no expected risks beyond those that may be associated with computer use.

What happens if I do not want to participate in this study or if I change my mind later?

Participation in this study is voluntary. You can decide for yourself whether or not you want to take part in this research. You can withdraw from this survey at any time without giving a reason, you will not be charged. If you decide to withdraw from the study, all data collected from you up to that point will be deleted.

How do the researchers protect my information?

The audio recording is stored on a secure server, after which it can be transcribed so that the opinions of the participants can be recorded in text. The audio recording is safely deleted after the transcription has taken place (within 4 months of the examination). The transcript is anonymised so it doesn't contain any information that could identify you. This anonymised transcript will be used in this research and possibly in later publications and studies.

Who has access to the information collected during this study?

Efforts will be made to limit the use and disclosure of your personal information, including research study reports, to people who need to review this information. We cannot promise complete secrecy.

There are reasons why information about you may be used or seen by other people outside the study during or after this study. University officials, auditors and Review Board may need access to the study information to ensure that the study is conducted in a safe and appropriate manner.

How can the information collected in this study be shared in the future?
We keep the information we collect about you during this research for study registration and for possible use in future research projects. Your name and other information that can directly identify you is stored securely and separated from the rest of the research information we collect from you.

Anonymised data from this research can be shared with the research community, with journals in which research results are published, and with databases and data repositories used for research. We will delete or encrypt any personal information that could directly identify you before the survey data is shared. Despite these measures, we cannot guarantee the anonymity of your personal data.

The results of this research may be shared in articles and presentations, but will not contain information that identifies you unless you allow for the use of information that identifies you in articles and presentations.

Do I get paid or do I get a contribution for participation in this study?
You will not receive any contributions for participation in this study.

Who can I talk to?
If you have any questions, comments or complaints, please contact the principal investigators, student researcher or the supporting institutions.

If you want a copy of this permission for your own administration, you can print it from the screen. If you are unable to complete this, you can contact the student researcher using the contact details above.

If you would like to participate, please tick the "I agree" box below and we will start the interview study.

If you do not wish to participate in this study, please check the box "I do not agree" below and close the browser.

○ I agree

○ I don’t agree

C. Questionnaire demographics

Q1 What is your gender?

○ Male
○ Woman
○ Non-binary
○ I'd rather not day
○ Other, namely: ...

Q2 What age range do you fall into?

○ 19 years or younger
○ 20-29
○ 30-39
○ 40-49
○ 50-59
○ 60 years or older

Q3 What is your highest level of training?

○ PhD
○ Master degree
○ Bachelor degree
○ High School
○ None

Q4 What is your specialised field of education (e.g., management, computer science, or no specialisation)?

Q5 What is your position within the police?

Q6 Can you briefly explain in 1 sentence what your activities are within your position?

Q7 How many years of experience do you have in this sector?

Q8 To what extent do you consider decision-making (choosing an action over alternative actions) as the primary task in your position? Choose the answer that best fits your job description.

○ Not at all. It is not my job to make decisions, nor to suggest recommended actions to decision-makers.

○ It is not my job to make decisions, but I suggest recommended actions to decision-makers.

○ I occasionally make decisions within my organisation, but do not do this on a regular basis.

○ I regularly make decisions within my organisation about a single department or organisational unit.

○ I regularly make decisions within my organisation about multiple departments or organisational units.

Q9 To what extent do you consider data analysis to be a primary task in your role? Choose the answer that best fits your job description.

○ It's not my job to analyse data, nor does it incorporate the work of data analysts.

○ It's not my job to analyse data, but I do incorporate the work of the data analysts to make decisions.

○ When necessary, I perform data analysis, but do not do this on a regular basis.

○ Performing data analyses regularly is my primary job.

Q10 How would you classify your knowledge about visualising data such as graphs (line graphs, bar graphs, ...), charts, etc.?

○ No knowledge: I have no knowledge of data visualisation.

○ Basic knowledge: I have a general knowledge and understanding of basis visualisation techniques and concepts (such as line graphs, bar graphs, etc.).

○ Beginner: As part of a training I learned about visualisations, but there is still a lot to learn for me.

○ Average: I am able to create visualisations and read them myself. I still need help creating and reading more complex visualisations.

○ Advanced: I effortlessly create and read visualisations.

○ Expert: I am a visualisation researcher or practitioner with a longer practice in the field of visualisations.

Q11 How would you classify your knowledge about Machine Learning?

○ No knowledge: I have no knowledge of Machine Learning.

○ Basic knowledge: I have a general knowledge and understanding of basic Machine Learning techniques and concepts.

○ Beginner: As part of a training I learned about Machine learning, but there is still a lot to learn for me.

○ Advanced: I effortlessly create and apply Machine Learning models.

○ Expert: I am a researcher or experiential expert in the Machine Learning field.

Q12 How would you classify your knowledge about Explainable Artificial Intelligence?

○ No knowledge: I have no knowledge of Explainable Artificial Intelligence.

○ Basic knowledge: I have general knowledge and understanding of basic Explainable Artificial Intelligence concepts.

○ Beginner: As part of a training I learned about Explainable Artificial Intelligence, but there is still a lot to learn for me.

○ Advanced: I effortlessly create and apply Explainable Artificial Intelligence models.

○ Expert: I am a researcher or experiential expert in the Explainable Artificial Intelligence field.

Q13 Here are some types of machine learning techniques that may be relevant to law enforcement. To what extent do you see potential in the use of the following techniques by the police?

I see potential in the technology...

	I do not know this technique	Totally disagree	disagree	neutral	agree	totally agree
Classification	○	○	○	○	○	○
Regression	○	○	○	○	○	○
Ranking & search	○	○	○	○	○	○
Semi-supervised learning (+ Active learning)	○	○	○	○	○	○
Dimensionality reduction	○	○	○	○	○	○
Clustering	○	○	○	○	○	○
Representation learning & Optimisation	○	○	○	○	○	○
Generative models & Virtual Reality	○	○	○	○	○	○
Reinforcement learning & Graph Analysis	○	○	○	○	○	○

Q14 Here are several areas of application that may be relevant to the police. To what extent do you find the application areas below relevant for the police?

I find the application area . . . relevant

	I am not familiar with application area	Totally disagree	disagree	neutral	agree	totally agree
Computer Vision (images & video)	○	○	○	○	○	○
Natural Language Processing (text)	○	○	○	○	○	○
Data Mining & Analysis (tabular)	○	○	○	○	○	○
Audio/Speech Recognition & Synthesis (audio)	○	○	○	○	○	○
Robotics / Intelligent Agents	○	○	○	○	○	○
Expert Systems / Knowledge Representation and Reasoning	○	○	○	○	○	○
Planning, Scheduling & Optimisation	○	○	○	○	○	○
Augmented & Virtual Reality	○	○	○	○	○	○
Social Networks & Graph Analysis	○	○	○	○	○	○

D. Interview protocol

Materials needed for the interview:

- ○ Laptop with the protocol (and for back-up recording)
- ○ Phone for audio recording
- ○ Papers
- ○ Pens
- ○ Water
- ○ Chargers (for phone, laptop, and tablet)
- ○ Tablet (for the online consent form and demographical questionnaire)
- ○ A table for quick notes

Interview introduction:

Welcome, my name is Kim and I will do the interview with you today. Thank you for being here. Before we start, I would like to ask you to read the consent form and decide if you agree to it.

Consent form: see Appendix B (during the interviews there was a link to the actual form)

[After they read the consent form]

During this interview we will discuss your experiences within law enforcement (police) and the types of decisions you make within your position. The purpose of the interview is for me to learn about the difficulties in your day-to-day decisions within your role and interactions with data so that I can understand how we can design better solutions to help you. After the interview, I remain at your disposal for any questions. Shall we start the recording now?

[After recording starts]

(if yes, after rec say)-> Do you confirm that you have read and agreed to the consent form to participate in this interview?

Short break to let participant answer.

As stated in the consent form, I remind you that this study is only registered for research purposes. Shall we start with the questions?

Beginning of the interview:

1. In this interview we want to discuss what decisions you need to make in your daily tasks/work and about what information you make these decisions and which technologies and tools you use in this process.

   Before we get into this, could you perhaps briefly explain to me what your tasks are within your position at the police? / What does a normal working day look like for you?

2. I want to ask you to remember a day when you had to accomplish a difficult task. Can you guide me through that day?

   - What steps have you taken to accomplish that difficult task?

   Since I have to work out the interview in detail and be precise in this, I would like to see what [tool XX] looks like and take screenshots of this.

   If they can't show it, ask:

3. What exactly are we looking at?

4. Could you describe in detail what we see here?

The example of the difficult task should include the following aspects:

1. The decision

   - What decision had to be made?

2. Information

   - What information did you use? / What information is relevant to your decision?
   - How did you get this information?
     - Did you have to search for this information?
     - Has anyone given you this information? If so,
       * Who gave you this information?
       * What is their role?
     - How is this information displayed? Can you show me what this information looks like?

3. Technology used

   - Have you used tools to help you with the task?
   - Do you understand this tool?
     - Why or why not?
     - If not: What could help to understand this tool?
   - What do the results of this tool look like? Can you show me?

4. Ohter people

   - Do you communicate the information you've gained with other people? If so,

  - Who are they?
  - What is their role?

  If not,

  - How do you use the results?/What do you do with these results?

- Are other people involved in the process?
- What kind of tools do you use to communicate the results?
  - Can you show me this one?
- How do these tools help to explain the results?

Demographics: see Appendix C (during the interviews there was a link to the actual form)

Conclusion of the interview:

We reached the end of the interview. Is there anything else you'd like to comment on? Or are there things that haven't been discussed yet, but that you think are important?

Then I would like to thank you very much for your time and participation.

Some sample questions to guide the discussion:

Decision	What kind of information do you use when making the decision?
	How did you get this information?
doubt	Why can't you explain how you obtained the information?
	Why can't you explain how to xxx?
	Why is it easy to follow?
	Why is it unclear?
Data	Can you give me more information on how to get xxx?
Reguide the interview	Earlier you said..... could you perhaps elaborate on that?
General probing questions	Can you elaborate on that?
	Can you give an example?
	What makes you think ...?
	What do you mean?
	So, am I right to conclude that....?
	Silence (tolerance)

E. Consent form

Title of research study: UX/UI design Explabox

Dear participant,

Before deciding to participate in this scientific study, carefully read the following information.

Principal Investigators: Marcel Robeer, marcel.robeer@politie.nl and Evanthia Dimara, e.dimara@uu.nl.

Student researcher: Kim van Genderen, m.s.vangenderen@students.uu.nl.

Supported by: This research is supported by the National Police Lab AI, nationaal-politielab@uu.nl.

Key study information:
Below is a brief summary to assist you in deciding about participation. More detailed information will follow.

- The purpose of this study is to find out if a specific interface design aids your decision making process.
- Anticipated research study participation is approximately 1 hour.
- Participation involves no greater risks than daily life and is comparable to typical computer usage.
- Based on a Quick Scan Ethics and Privacy, this research has received approval from the Research Institute of Information and Computing Sciences. If you have a complaint or comment about the way this investigation was conducted, you can send an email to ics-ethics@uu.nl. For questions or complaints about the processing of personal data, you can send an e-mailto the Data Protection Officer of Utrecht University via privacy-beta@uu.nl.

Why am I being asked to participate?
We ask you to participate in this investigation because you are between 18 and 65 years old, and are involved in decision-making as part of your position within the police.

What should I know about participating in a research study?

- Participation in this study is entirely voluntary, whether or not you take part is up to you.
- You can choose not to take part.
- You can agree to take part and later change your mind.
- Your decision will not be held against you.
- You can ask all the questions you want before you decide.
- You do not have to answer any question you do not want to answer.
- You can quit anytime during the study. This can be done after collecting the data, but also during the collection of the data.

What happens if I say, "Yes, I want to participate in this study"?

- The evaluation will be recorded (audio). You agree to this if you consent to participate in the research. The recordings will be used for research purposes only.
- During the first part of the evaluation session, some short general questions are asked.
- The second part deals with the designed interface.

What happens if I do not want to be in this research, or I change my mind later?
Participation in this research is voluntary. You can decide for yourself whether or not you want to participate in this study. You can withdraw from this study at any time without giving a reason, it will not be held against you. If you decide to withdraw from the study, all data collected from you up to that point will be deleted.

How will the researchers protect my information?
The audio recording is stored on a secure server, after which it can be transcribed so that the opinions of the participants can be captured in text. The audio recording will be safely deleted after the transcription has taken place (within 4 months after the study). The transcript is anonymised so that it does not contain any information that could identify you. This anonymised transcript will be used in this research and possibly in subsequent publications and studies.

Who will have access to the information collected during this research study?
Efforts will be made to limit the use and disclosure of your personal information, including research study records, to people who have a need to review this information. We cannot promise complete secrecy.

There are reasons why information about you may be used or seen by other people beyond the research team during or after this study. Examples include:

- University officials, auditors, and Review Board may need access to the study information to make sure the study is done in a safe and appropriate manner.

How might the information collected in this study be shared in the future? We will keep the information we collect about you during this research study for study recordkeeping and for potential use in future research projects. Your name and other information that can directly identify you will be stored securely and separately from the rest of the research information we collect from you.

De-identified data from this study may be shared with the research community, with journals in which study results are published, and with databases and data repositories used for research. We will remove or code any personal information that could directly identify you before the study data are shared. Despite these measures, we cannot guarantee the anonymity of your personal data.

The results of this study could be shared in articles and presentations, but will not include any information that identifies you unless you give permission for use of information that identifies you in articles and presentations.

Will I be paid or given anything for taking part in this study?
You will not receive compensation for participating in this study.

Who can I talk to?
If you have any questions, concerns or complaints, please contact the Principal Investigators, Student Researcher or the Supporting Institutions.

If you would like a copy of this consent for your own records, you can print it from the screen. If you cannot print it, you can contact the Student Investigator with the contact information above.

If you would like to participate, please tick the "I agree" box below and we will start the interview study.

If you do not wish to participate in this study, please check the box "I do not agree" below and close the browser.

○ I agree

○ I don’t agree

F. Evaluation protocol

Evaluation introduction:
Welcome! I'm Kim, and I'll be guiding today's evaluation. Thank you for participating. Your insights and opinions are important, and there are no wrong answers.

Consent form: see Appendix E (during the evaluation sessions there was a link to the actual form)

[After they read the consent form, start recording]

Let the participants verbally agree that they gave consent

Next, I will show you a tool about October events in Utrecht, in four categories: 1. events,2. demonstrations, 3. football, and 4. public order permit (O.O.V.). After a quick training, you'll try some tasks to show if you understand the tool.

[Get tool ready]

Here is the updated police calendar. The initial screen shows this week's events, categorised by four colors in the right-side legend: blue for demonstrations, red for events, purple for football, and green for public order permits (O.O.V.). Hover over a color, you will see three risk classes: lighter shades indicate lower risk, and it darkens as risk increases.

Events are rated by a human (person icon) and machine (robot icon), each indicating their confidence level in the assessment. In the legend: no lines mean high confidence, close lines indicate uncertainty, and mid-spaced lines signify moderate confidence.

Beginning of the interview:

- Could you now click on a football event from the current week?
- Could you now click on an event from the current week?
- Could you now click on a demonstration from within the current week?

[IF CORRECT CLICKS MADE, PROCEED. OTHERWISE, TRY AGAIN]

The screen provides detailed information about the demonstration. On the left, you'll see data used by the human for assessment and on the right, data used by the machine for assessment.

The machine figures out risk by giving scores to certain words and adding them up for a total risk score.

Clicking back takes you to the weekly overview.

Can you now click on an event within the current week and tell me what the total score is?
And, do the same for an O.O.V. event within this week?

[IF TWO TOTAL SCORES ARE CORRECTLY SHOWN, CONTINUE]

Navigate through October by clicking the arrows atop the calendar. Can you show me events from one weeks ago.
And could you show me the events that will take place in two weeks?

Now, click on the month button to view all October events.
Please click a low-risk demonstration, according to the machine.

Return to the weekly calendar and click the map button.

This displays daily Utrecht events on a map. Can you find and select the demonstration on Friday, October 20, and tell me its location?

Training is complete; let's begin the tasks.

[Think-aloud protocol with some tasks]

I will give you the tasks one by one. one at a time. While you perform each task, please think out loud, sharing your thoughts and actions. Once a task is complete, I'll provide the next one.

1. Can you select an incident that takes place on Wednesday 18th?
   (a) Where does this event incident take place?
   (b) How many visitors are expected?
   (c) What is the risk classification according to the machine?
   (d) What is the risk classification according to the human?
2. Can you tell me on which day in October most events take place?
3. Can you select an event within the week from October 9 to 15, for which the risk classification is different for the human and the machine? This can be any type of event. However the human needs to have given a classification.
   (a) Could you tell me what risk classification the machine gives this event, and which classification the human gives for this event?
4. Could you select a demonstration for which the machine is the (most) uncertain about, in the week from October 16 to 22?
   (a) Can you give me the exact percentage of the machine's certainty score?
   (b) Can you tell me what risk rating the machine gave to this demonstration?

   (c) Can you also tell me why you think the machine gives this risk indication classification?

5. Can you click on an event ("gebeurtenis") during the week from October 23 to 29. that has an assessment of the machine but no assessment of the human? This can be of any type of event.
   (a) For this event, would you trust the machine or would you still want the human to look at it?
   (b) Why?

[DECISION TASK]

Imagine you are responsible for allocating police resources in the last week of October. Using the new dataset, please identify which event, from any category, will require the most police resources and explain why. Be sure to speak aloud all your thoughts, considerations, and the information you use from the tool to make your decision. Every bit of your thought process is valuable here!

[GIVE NEW DATASET] [SOME FINAL QUESTIONS I WOULD LIKE TO ADD TO THE CURRENT SURVEY]

Next, I'd like to ask you some additional questions about your experience with the tool, focusing on both the positive aspects and potential areas for improvement . Specifically:

Helpful Aspects of the Tool:

- What parts of the tool helped you make decisions?
- Can you explain why these parts were helpful?

Comparison with Current Police Planning Design:

- Comparing this tool design with the current version for police planning, do you think it assists you better or worse in planning and deciding how to allocate resources?
- Why do you feel that way?

Challenges with the Tool:

- Were there aspects of the tool that were confusing or didn't help much when you were making decisions?
- Why do you think these aspects were challenging or unhelpful?

Unmet Needs or Missing Features:

- Did you find yourself wishing the tool could do something that it did not?
- What additional features or information would have helped you make your decision?

Enhancements for Existing Features:

- Was there a feature that you found somewhat useful but think could be improved or expanded upon?
- In what way could this feature be improved to better assist you?

Final Question:

- Do you have any other reflections, whether positive or negative, that you'd like to share about your experience using the tool?

[DEMOGRAPHIC SURVEY]