



Tim Wei Shi Vrieling

Using Explanations to Foster Values: Expert Opinions on the Potential of Broad Explainability to Foster Different Values in the Context of Predictive Policing

Master Thesis

at the Chair for Information Systems and Information Management
(Westfälische Wilhelms-Universität, Münster)

Supervisor: Prof. Anu Masso
Co-supervisor: Colin van Noordt

Presented by: Tim Wei Shi Vrieling
Stoepveldsingel 95
9403SM, Assen, Netherlands
+31618073292
tvrielin@uni-muenster.de

Date of Submission: 2021-08-09

Content

Abstract	IV
Tables	V
Abbreviations	VI
1 Introduction	1
2 Research Background	4
2.1 Researching the Use of Algorithms in the Public Sector	4
2.2 Predictive Policing: A Holistic View	6
2.3 Important Values in the Context of Predictive Policing	11
2.3.1 The Main Benefits of Predictive Policing	11
2.3.2 Main Concerns regarding Predictive Policing	13
2.4 Explainability: A Broader Perspective	19
2.5 Fostering Values with Broad Explanations	24
2.5.1 Determining the Content of Explanations	26
2.5.2 Algorithmization: Aspects of the Socio-Technical Context	29
3 Research Design and Methodology	31
3.1 Research Design	31
3.2 Methodology	32
3.2.1 Exploratory Expert Interviews	33
3.2.2 Sampling Method and Expert Sample	36
3.3 Data Analysis Method	39
4 Results	42
4.1 General Comments on Explainability	42
4.2 Results for each Individual Value	44
4.2.1 Values Resulting from the Benefits of Predictive Policing	44
4.2.1.1 <i>Efficiency</i>	44
4.2.1.2 <i>Effectiveness</i>	47
4.2.1.3 <i>Accuracy</i>	49
4.2.1.4 <i>Security</i>	51
4.2.2 Values Resulting from Concerns with Predictive Policing	53
4.2.2.1 <i>Explainability</i>	53
4.2.2.2 <i>Accountability</i>	55
4.2.2.3 <i>Responsibility</i>	58
4.2.2.4 <i>Transparency</i>	59
4.2.2.5 <i>Comprehensibility</i>	62
4.2.2.6 <i>Trust</i>	63
4.2.2.7 <i>Fairness</i>	66
4.2.2.8 <i>Privacy</i>	68
5 Discussion	70
5.1 Interpretation of the Results	70
5.1.1 Broad Explainability and Other Values: Fostering or Balancing?	70
5.1.2 How Broad Explanations can Foster Values: Factors of Explanation	77
5.1.2.1 <i>Audience</i>	77
5.1.2.2 <i>Related Values</i>	79
5.1.2.3 <i>Components of Algorithmization</i>	81
5.2 Theoretical and Practical Implications	83

5.3 Limitations and Future Research.....	85
6 Conclusion.....	87
References	89
Appendix	91

Abstract

This thesis aims to address gaps in our current knowledge and understanding with regards to explainability in a practical context, and conduct research on how explainability interacts with other values and how giving explanations could actually foster particular values in said practical context. To address this aim, this thesis seeks to answer the following research question: *How would experts describe the interaction between broad explainability and other important values in the context of predictive policing, and how, according to them, could explanations foster these values?* To answer this research question a systemic literature review was conducted on the basis of which several theoretical and conceptual elements were identified which formed the basis for data collection and analysis. First, 12 values were identified, after which, the concept of broad explainability – *the act of giving explanations that are aimed at explaining multiple aspects of the socio-technical context* – was defined. Additionally, two ways in which broad explainability could relate to predictive policing were outlined: fostering and balancing, as well as, three factors of explanation – audience, related values, and content. Finally, on the topic of content, 7 components of algorithmization along, which the content of broad explanations could be structured, were also described. This thesis followed a qualitative research design, and used an exploratory expert interview methodology to answer the research question. The interviews were structured in two parts, each pertaining to a certain part of the research question. They were recorded, transcribed, and subsequently coded and analysed. In the end, 6 main findings are outlined which contribute to both the existing literature and existing practice with regards to explainability.

Tables

Table 2.1: Important Values in the Context of Predictive Policing	24
Table 2.2: Types of Transparency	26
Table 4.1: Main Takeaways – Efficiency	47
Table 4.2: Main Takeaways – Effectiveness	49
Table 4.3: Main Takeaways – Accuracy	51
Table 4.4: Main Takeaways – Security	53
Table 4.5: Main Takeaways – Explainability	55
Table 4.6: Main Takeaways – Accountability	57
Table 4.7: Main Takeaways – Responsibility	59
Table 4.8: Main Takeaways – Transparency	61
Table 4.9: Main Takeaways – Comprehensibility	63
Table 4.10: Main Takeaways – Trust	65
Table 4.11: Main Takeaways – Fairness	68
Table 4.12: Main Takeaways – Privacy	69
Table 5.1: Categories of Audience	78
Table 5.2: Important Values and their Related Values	80
Table 5.3: Components of Algorithmization needed to Foster Values	82

Abbreviations

AI	Artificial Intelligence
XAI	Explainable AI

1 Introduction

This thesis aims to address gaps in our current knowledge and understanding with regards to explainability in a practical context, and conduct research on how explainability interacts with other values and how giving explanations could actually foster particular values in said practical context. In recent years, the public sector has been using a number of different types of algorithms to support an increasingly broad range of tasks. For example, algorithms are used in education, to find and select the most effective teacher when recruiting (Rockoff, Jacob, Kane & Staiger, 2011). Or in criminal law, where they are used by judiciaries to determine the risk that a criminal might re-offend after making bail (Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018). Predictive algorithms are one of these types of algorithms that have experienced an uptake by public sector organisations in recent years. The use of predictive algorithms, also referred to as predictive analytics, revolves around the creation of models based on historical and current data, to provide predictions about future behaviour or events (De Laat, 2019; Ogunleye, 2014).

Because predictive algorithms are used to make sense of large complex Big Data sets, they are often very complex themselves (Ogunleye, 2014). This has exacerbated the so called black-box problem. As Rai (2019) explains, newer, more complex forms of algorithms, such as those based on machine learning or deep neural networks, sacrifice transparency and interpretability for accuracy of predictions, sometimes to the extent that developers themselves also cannot understand how the algorithms produce decisions anymore (Bertossi & Geerts, 2020; Samek & Müller, 2019).

This is especially problematic for the public sector because, generally speaking, predictive algorithms are increasingly used to either replace or support human decision making (Zarsky, 2016). For example in healthcare, where, they are used to predict, e.g. the risk of patients having a certain disease (Henriksen & Bechmann, 2020), or in social welfare where they are used to predict the risk that a child is mistreated or the risk of recurring domestic violence (Gillingham, 2019). But due to their black-box nature, issues are arising with regards to trust (Tene & Polonetsky, 2017), fairness (Corbett-Davies et al., 2017) and accountability (Veale, van Kleek, & Binns, 2018) among others.

As an answer to these black-box problems, explainability and explainable AI (XAI) have been suggested by a multitude of scholars who study complex, algorithm based, systems (Miller, 2019; Rai, 2019; Bertossi & Geerts, 2020; Hansen & Rieger,

2019). Furthermore, explainability is also increasingly mentioned in guidelines and principles on AI (Hagendorff, 2020). Explainability as a value can be understood as: *the extent to which information can be communicated in a way that can be understood by recipients* (Samek & Müller, 2019; Rai, 2019; Meijer & Grimmelikhuijsen, 2020). The benefit of explainability would be that it would provide a basis for transparency, justification and traceability which could foster trust in black-box systems (Shaban-Nejad, Michalowski, & Buckeridge, 2021). Some authors even go so far as to link explainability directly to increases in trust, transparency and fairness (Hansen & Rieger, 2019; Shaban-Nejad, Michalowski, & Buckeridge, 2021; Samek & Müller, 2019; Weller, 2019).

Methods of realizing explainability, so called XAI methods, are mainly targeted at explaining how an algorithm produces outputs (Rai, 2019; Samek, Montavon, Vedaldi, Hansen and Müller, 2019). As Hagendorff (2020) explains, explainability is mostly implemented by means of technological solutions. However, this thesis will argue that these technical explanations are not enough to foster values, because they might be incomprehensible or irrelevant. As Miller (2019) explains it, the developers who develop XAI methods do not seem to take into account whether lay people would actually be able to understand the explanations that are produced with by these methods. Furthermore, this thesis will add that technology-centred explanations might be irrelevant to fostering values if we consider the larger socio-technical context.

Besides these technology-centred XAI explanations, there are no other suggestions in literature on how explainability could be practiced and how this practice could foster value. Most authors who argue for explainability seem to assume that being able to explain how an algorithm produces a result will automatically foster values such as accountability, transparency, trust and fairness (Samek & Müller, 2019; Sandhu & Fussey (2020); Haque, Weathington, Chudzik and Guha, 2020; Rai, 2019). So there is a clear gap in our current knowledge on how explainability could actually foster different values. A gap that needs to be filled because many scholars and practitioners are viewing explainability as the new silver bullet, while our current methods of practicing explainability seem to be inadequate to actually fulfil the promise of explainability: fostering important values.

The specific empirical aims of this thesis are therefore twofold: (1) analyse how explainability interacts with other values in a practical context, and (2) analyse how explanations could actually foster certain values. Through analysing this, this thesis also

aims to contribute to the prior academic discussions about the conceptual understanding of explainability, by arguing that a broader approach needs to be taken in order to foster values. This thesis will consider explainability in the context of predictive policing – which refers to the police’s use of predictive algorithms to pre-empt and prevent crime (Hälterlein, 2021). This context was chosen because it is one of the most prominent places where predictive algorithms are seeing increased application (Meijer & Wessels, 2019), and because this is one of the main areas in which the use of algorithms is causing concerns (Szczepanski¹, Pawlicki, & Pawlicka, 2021).

In order to enable research into the chosen topic, and provide a basis for answer the research question, a systemic literature review was conducted following the methods of Webster and Watson (2002). Based on this, 12 important values in the context of predictive policing were identified in this thesis. After which, the concept of broad explainability - *the act of giving explanations that are aimed at explaining multiple aspects of the socio-technical context* – was defined. Then, two ways in which broad explainability could relate to predictive policing were also defined: fostering and balancing. As well as, three factors of explanation – goal, audience, and content. Ending with a description of 7 components of algorithmization along which the content of broad explanations could be structure. These elements then formed the theoretical and conceptual basis to answering the following research question:

How would experts describe the interaction between broad explainability and other important values in the context of predictive policing, and how, according to them, could explanations foster these values?

This thesis will follow a qualitative research design, and use an exploratory expert interview methodology to answer the research question. The interviews were structured in two parts, each pertaining to a certain part of the research question. They were recorded, transcribed, and subsequently coded and analysed.

The thesis is structured as follows, the next part will outline the research background and conclude with having outlined and identified the aforementioned elements. After this the research design and methodology will be discussed, followed by the results of the interviews. Then, the results will be discussed and 6 main findings will be outlined. This section will also include a discussion of the theoretical and practical implications of the main findings, the limitations and suggestions for further research. The thesis will end with a general conclusion.

2 Research Background

2.1 Researching the Use of Algorithms in the Public Sector

The public sector has been using algorithms to support an increasingly broad range of tasks. For example, algorithms are used in education, to find and select the most effective teacher when recruiting (Rockoff, Jacob, Kane & Staiger, 2011); they are used by police to predict crime patterns (Meijer & Wessels, 2019); and they are used by judiciaries to determine the risk that a criminal might re-offend after making bail (Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018). Generally speaking they are increasingly used to either replace or support human decision making (Zarsky, 2016). Between these different examples, however, several different technologies seem to be under discussion. What are the differences between these technologies? Or are they the same? Often these technologies are grouped together under the term Artificial Intelligence (AI), but a systemic literature review on artificial intelligence research points to the fact that there are huge ambiguities with regards to the definition of AI (Collins, Dennehy, Conboy, & Mikalef, 2021).

Thierer, O'Sullivan, & Russell (2017), give a clear overview of what they understand when talking about AI and related concepts. Based on their definitions, it becomes clear that algorithms could be considered a part of AI, specifically, the enabling part. Algorithms determine the operations that need to be carried out in order to perform a task. It is at the point when these tasks are executed well enough that the machine/system appears to be intelligent, that we talk about AI instead of mere algorithms. Clearly, it is the idea of intelligence portrayed by a machine or system that sets AI apart from algorithms. But what is intelligence? And when does a machine or system behave intelligently? As Miailhe and Hodes (2017) point out, experts have been unsuccessful in outlining the limits of AI because they have been unable to come up with a universally accepted definition of intelligence.

Several approaches have been taken towards conceptualizing what intelligence is and when a machine or system portrays it. As Russel, Norvig, and Davis (2016) explain – attempts to define the intelligence of a machine can be split into four categories: thinking humanly, acting humanly, thinking rationally, and acting rationally. However, each of these attempts brings forth its own ambiguities as to what specifically would be considered 'intelligent' (thinking or behaviour) and what is the root of intelligence

(human nature or rationality). Also the question of how to measure this in a machine remains unanswered.

Another concept which lies at the intersection of the conceptual ambiguity surrounding AI is ‘machine learning’, which implies an additional dimension to the idea of machine intelligence – the ability to learn. Thierer, O’Sullivan, & Russell (2017), understand machine learning as “the process by which a computer can train and improve an algorithm or model without step-by-step human involvement (p.9).” This definition shows that algorithms, through autonomous training, can be improved. Which can also be interpreted as: ‘taught to behave more intelligent’. This definition shows how machine learning links to AI and algorithms. Where algorithms are a series of instructions that make computers perform tasks producing desired results, AI is the name that is given to a machine when they perform those tasks so well that it mimics intelligence; whatever that may entail. Machine learning, is the process by which a seemingly unintelligent algorithm can autonomously learn to be more intelligent. Thus constituting a method of turning simple algorithms into AI.

It may now be clear that algorithms, AI, and machine learning are all very closely related, which causes a lot of ambiguity among researchers with regards to how to define AI (Collins, Dennehy, Conboy, & Mikalef, 2021). On top of this, there is the ‘AI Effect’: a phenomenon where, as soon as some machine or system approaches the ideal of machine intelligence, it gets scrutinized to the point that it cannot be considered ‘true AI’ (OECD, 2019); or in other words, not truly intelligent. As long as there is no clarity as to what machine intelligence is, how it can be measured, and the AI Effect continues, it becomes very hard to be concrete when writing about AI in the context of algorithm-based innovations.

In conclusion, a wide variety of algorithm based technologies, displaying varying degrees and forms of intelligence, are referred to as AI, and experts are still not certain about the exact limits of the concept of AI. Therefore, while conducting research into these technologies, this thesis adopts the stance that it would be better to talk about a specific instance of algorithm use, within a specific field, as the subject of study instead of claiming to study AI in general. It is therefore, that this thesis will narrow its focus down to cover specifically – predictive algorithms and their use by police organizations. This however, still leaves a broad range of different algorithm based technologies within the scope of this thesis, as shall be showcased below. Which is another example of how broad and varied the different technologies are that fall under the umbrella of AI. On the

flip side, this also means that the results gained from studying this specific instance of algorithm use by the public sector are still relevant in the broader context of 'AI' research.

2.2 Predictive Policing: A Holistic View

In conjunction with the rise of Big Data came the expansion of state surveillance in the western world as a reaction to 9/11. As a result of these two trends coinciding, governments started to look at the possibilities of using Big Data analytics for police surveillance (Brayne, 2017). Especially in the United States predictive algorithms are finding application within police departments (Bayrne, 2017; Bennet Moses & Chan, 2018), and in other countries such as China, Denmark, Germany, India, the Netherlands, and the United Kingdom, predictive policing tools are finding uses at the local level (McCarthy, 2019). Hälterlein (2021) explains that predictive policing is concerned with algorithmic crime forecasts. Put more specifically, predictive policing “makes use of crime forecasts based on a predictive model relying on multivariate methods that use current and past values of independent variables to predict the future value of the dependent variable (Hälterlein, 2021, p.3).” As such, we can understand predictive policing as the practice of using predictive models to forecast future crime, the aim of which, according to Hälterlein (2021), is to prevent this crime. This is also in line with Chan & Bennet Moses (2016), according to whom predictive policing can be regarded as a form of pre-emptive policing using statistical data.

Ferguson (2017), similarly understands crime prevention to be the goal of predictive policing, and distinguishes between two types: (1) Person-based predictive policing; (2) Place-based predictive policing. As the names imply, the first instance uses predictive models to identify potential criminals and potential victims. The second instance uses these models to identify crime patterns and ties those to geographical locations in order to predict where crimes are likely to take place. Comparing this to Hälterlein's (2021) explanation of predictive policing however, we see that Ferguson's (2017) distinction only covers one differentiating factor of predictive policing. Besides person- and place-based predictive policing, there are other factors that must be considered. What data is processed, what type of crimes are predicted, and whether patterns are sought in environmental factors or behaviour, all help determine what predictive model is created and as such, have impact on the predictions that are produced and the way future crime is governed (Hälterlein, 2021).

Hälterlein (2021), outlines three epistemologies of predictive policing that differ in terms of:

1. The relevance that is ascribed to the subject-matter theories;
2. The limits of predictions that are set;
3. The general explanations of crime that are given; and
4. The conditions for implementing algorithmic accountability (Hälterlein, 2021, p.2).

Based on these differences, Hälterlein (2021) argues that employing different epistemologies of predictive policing can lead to major differences in how future crimes are made knowledgeable and what action is taken based on this knowledge. Furthermore, it also has significant implications for the status of criminal knowledge in the justice system, where the attention of law enforcement is directed, whether there is a need for predictions to be 'meaningful' and finally, whether professionals can understand the algorithmic systems or not.

It is important to take note of these differentiating factors of predictive policing and their impact, because it shows that there is not a single approach to predictive policing but rather a broad variety of approaches depending on the chosen epistemology. However, the most important thing to take away from Hälterlein's (2021) paper, is that human decisions are still at the heart of policing even if a large part of it is based on algorithmic predictions. As Hälterlein (2021) explains, each of the epistemologies discussed represents a distinct way in which a predictive model can be constructed and specified through an exercise called 'parametrisation' – "the process of choosing independent variables that represent relevant aspects of the subject-matter problem (p.3)." As such, the differences in predictive policing approaches stem from differences in human decisions with regards to the parameters of a predictive model, and because of this, it is a collection of human decisions that eventually determine how the picture of future crime is painted.

This thesis shall not further discuss the differences between predictive policing approaches as it is interested in studying broad explainability in the context of predictive policing understood from a holistic point of view – which implies across these different approaches. As such, when referring to predictive policing practice, it is to be understood that this refers to all instances in which predictive models are used to predict and prevent future crime regardless of the specific underlying predictive model. The downside of this

choice is that the significance of these differences might be overlooked. But it is necessary to be able to study how broad explainability as a practice interacts with other important values in the context of predictive policing.

So what is this holistic view on predictive policing? Let's start with explaining what the main impact of predictive policing is on police practice. The main impact is that it turns the focus of policing from being reactive to pre-emptive. This seems logical when considering the goal of predictive policing – preventing crime. Pre-emptive policing, as such, is based on the idea that police can take action even before a crime takes place (Van Brakel, & De Hert, 2011), and it can be assumed that moving from a reactionary to a pre-emptive/preventive method of crime fighting will require some drastic changes in the way the police operates. Therefore, there should be some significant benefits to predictive policing that motivate police departments to start using these technologies. Of course, the main benefit and ultimate goal of using predictive policing is a reduction in crime rates (Bennet Moses & Chan, 2018). But more specific claims on the potential benefits mostly boil down to improvements in the efficiency and accuracy of police actions and as a result – their effectiveness (Meijer & Wessels, 2019; Ferguson, 2017). Some however, also go so far as to claim that predictive policing can help improve accountability and combat discrimination (Ferguson, 2017; Bayrne, 2017).

What is strange however, is that there is often little proof provided to substantiate the claims that predictive policing helps reduce crime rates (Bennet Moses & Chan, 2018). On top of that, empirical research into the effectiveness of predictive policing is inconclusive or reluctant to definitively link the successful reduction in crime to just the use of predictive policing alone. To illustrate, Hunt, Saunders & Hollywood (2014) evaluated the Shreveport predictive policing experiment. In this experiment, three districts that were making use of predictive policing were compared with a control group of three districts that were not using these technologies. The evaluation concluded that there was no statistically significant difference between the districts using predictive policing and the control group. Their authors provided three possible reasons for this outcome: the study design was not comprehensive enough, the differences in the implementation of the programme were too significant across the districts to yield a suitable comparison, and/or the design of the programme was inadequate for producing crime reductions.

Mohler et al. (2015) on the other hand, found that Los Angeles police divisions using their predictive policing tool averaged a crime volume reduction of 7,4% because of more

effective patrols. Furthermore, more recently, Levine, Tish, Tasso, and Joy (2017) evaluated the use of predictive policing by the New York Police Department and they found a 6% crime reduction after implementation of their predictive policing technology. However, Mohler et al (2015), also mentioned that the generalisability of these results are heavily dependent on the extent to which the policing practices of these police divisions are representative of policing practices at other police divisions. Additionally, it was also acknowledged that predictive policing is effective only in the short term as a tool for disrupting crime opportunities but that, in order to be effective in the long term, there would be a need for crime reduction strategies aimed at the fundamental causes of crime. In similar fashion, Levine, Tish, Tasso, and Joy (2017) had to acknowledge the 6% crime reduction could not fully be attributed to the use of the predictive policing system alone.

The idea that a predictive policing system in itself is not a solution but should rather be considered as part of a larger crime fighting strategy is in line with earlier work from Santos (2014). She concluded that there was not enough evidence to definitively say that predictive policing was effective. She also suggested that crime analysis in itself cannot be said to reduce crime because there is no proven direct link between this practice and reductions in crime rates. Finally, she argued that crime analysis should be viewed, not as a solution in and of itself, but rather as a component of a more comprehensive crime reduction strategy. Taking this stance, she did go on to show that crime analysis is a key component in successful crime reduction efforts.

What becomes clear from this discussion about the benefits of predictive policing is that these benefits do not materialise merely from the implementation of these technologies, if they even materialise at all. Predictive policing has to be seen as part of a larger crime fighting strategy, and as such, it can be argued that they should be studied from a broader perspective than one that is only focussed on the technology. This argument is further supported by Bennet Moses and Chan (2018), who argue that, in order to fully determine whether a decision was made properly, not only the decision itself should be under scrutiny, or the data on the basis of which the decision was made, but the entire process – from data collection and storage to its use in the algorithm and the eventual decision output and action. In other words, when studying predictive policing, it is better to focus on the entire process, rather than only the technology.

To further elaborate on the argument for taking a more holistic view on predictive policing, earlier in this section, it was established that human decisions play a large role in the creation of a predictive model. Furthermore, as has been shown, authors are arguing

in favour of taking a more holistic view on predictive policing, as it has the potential to provide a better explanation of how predictive policing could be deployed beneficially (Mohler et al, 2015; Santos, 2014; Bennet Moses & Chan, 2018). Following this same logic, adopting such a holistic view should also have the potential to provide a better explanation when it comes to how broad explainability interacts with other important values in the context of predictive policing.

The idea that predictive policing technology should be considered as part of a larger crime fighting strategy, which also involves human decisions, is in essence an application of the socio-technical perspective to predictive policing. Scholars who take this perspective argue that technology does not operate in a vacuum but is part of a larger socio-technical system/context where humans interact with the technology in order to complete certain tasks and obtain certain objectives. This larger system itself should then also be considered as existing in a dynamic environment which necessitates a certain amount of flexibility in order to adjust to changes in said environment (Meguire, 2014).

Applying this perspective to predictive policing, it can be said that predictive policing is a socio-technical system in which the predictive algorithm provides predictions that allow police officers to pre-empt and prevent crime as part of a larger crime fighting strategy, which, in turn, has to adapt to the changing reality of crime. This perspective will form the basis of the holistic view on predictive policing adopted in this thesis. In other words, this thesis will not only consider the predictive policing algorithm, but also the activities surrounding the predictive algorithm such as the data collection & selection, the creation of predictions, actions taken based on the predictions, and the scrutinizing of these actions. Not forgetting the human decisions related to these activities. When studying how broad explainability interacts with other values in the context of predictive policing, taking this perspective is important.

This will become more apparent when discussing the important values in the context of predictive policing that this thesis will focus on. These values shall be identified and defined by means of a discussion of the main benefits of predictive policing and the main concerns with predictive policing. This method was chosen because this will also explain what these values actually mean in practice, which will result in definitions that are more suitable for a discussion on how broad explainability relates to these values as opposed to providing general definitions based on what is written in, for example, in AI guidelines or principles.

2.3 Important Values in the Context of Predictive Policing

2.3.1 The Main Benefits of Predictive Policing

As was briefly mentioned in the previous section, the main benefits of predictive policing are increases in effectiveness, efficiency and accuracy with the eventual goal of reducing crime rates (Bennet Moses & Chan, 2018; Ferguson 2017). While some even claim that predictive policing can help improve accountability and combat discrimination (Ferguson, 2017; Bayrne, 2017). This provides a starting point for defining which values are important in the context of predictive policing.

In the context of predictive policing, the gains in efficiency, accuracy, and effectiveness are often explained as related. For example, by Ferguson (2017), who explains that, the drastic improvements in Big Data driven technologies promises greater accuracy through the ability to analyse and compare much larger amounts of data. Greater accuracy means better predictions. This is where accuracy links to efficiency, as better predictions will enable the police to allocate resources better in both a temporal and geographical sense, leading to a reduction in efforts wasted. Hence the claim that predictive policing will improve the efficiency of policing (Meijer & Wessels, 2019; Ferguson, 2017). Finally, it is easy to see how these improvements in accuracy and efficiency can lead to more effectiveness. Predictive policing could enable police departments to anticipate better where to deploy their resources as well as who to target and who to protect – which in turn should lead to more effective policing operations exemplified by lower crime rates.

Based on this discussion it is clear that efficiency, effectiveness and accuracy, can be considered as important values in the context of predictive policing. They can be defined as follows: (1) Accuracy, meaning *the extent to which a predictive algorithm makes correct predictions*, (2) Efficiency, meaning *the extent to which police resources are optimally allocated in spatial and temporal sense*, and (3) Effectiveness, meaning *the extent to which police operations are successfully reducing crime rates*.

Accuracy, efficiency and effectiveness, are very different from the types of values that are often represented in guidelines and principles on AI. These type of documents often discuss much more abstract values such as trust, fairness, explainability and privacy (Hagendorff, 2020). However, after conducting an extensive review of 22 major guidelines on AI ethics, Hagendorff (2020) remarked that many of these guidelines lack a consideration for “wider contexts and the comprehensive relationship networks in

which technical systems are embedded (p.103)”. In essence addressing the need to consider the socio-technical context in which the technology is used. This goes to show that identifying accuracy, efficiency and effectiveness as important values in the context of predictive policing is not a useless exercise, but might actually be necessary in order to get a more complete picture of how broad explainability as a practice might interact with these and other values in the larger socio-technical context of predictive policing.

Besides the claims surrounding efficiency, effectiveness and accuracy, there are also authors who argue that these technologies will help increase the transparency & accountability of policing practices and help combat discrimination (Ferguson, 2017; Bayrne, 2017). These claims are made on the basis of the assumption that the data on which predictions, and subsequently decisions, are based, can be easily provided in courts for evaluation, resulting in increased transparency of policing practices. Supposedly, this also improves the accountability of police departments (Ferguson, 2017). Bayrne (2017), goes a step further and argues that by removing a part of the human element in policing and replacing it with an allegedly unbiased algorithm, predictive policing can function as an antidote against discriminatory practices in police departments. However, these claims have even less empirical evidence than the supposed improvements in efficiency, accuracy and effectiveness (Meijer & Wessels, 2019), and a number of studies also argue for the potential of predictive policing to have the reverse effect and create other, more complex, accountability problems, decrease transparency and exacerbate discrimination (Bennet Moses & Chan, 2018; Ferguson, 2017; Brakel, 2016). As such, these values are better discussed in the next section which deals with the main concerns regarding predictive policing.

Looking at predictive policing from another point of view are Kasapoglu and Masso (2021), who studied security algorithms. According to them, security algorithms are algorithms that “ensure the safety of society (p.2).” With this they hint at another important value in the context of predictive policing: security. Security algorithms are essentially predictive policing algorithms as the term also refers to algorithms that are used to assess risk and predict future crimes with the goal of crime prevention (Hardyns and Rummens, 2018). However by using a different name than predictive policing algorithm to describe these types of algorithms, Kasapoglu and Masso (2021) put more focus on the ‘assessing risk’ aspect of predictive policing, rather than the ‘predicting and preventing crime’ aspect, like the term ‘predictive policing’ does. Focussing on police risk-scoring algorithms as a case, they studied how the understanding of security shifts in

the context of ‘security constructed through algorithms’. They did so by comparing the perspectives of data experts with refugees looking at both Estonia and Turkey. They concluded that refugees and data experts took opposing positions on security algorithms based on their conception of security. Differences were also found between refugees living in Turkey and in Estonia.

The research by Kasapoglu and Masso (2021) shows that, the perception and understanding of security differs based on who is asked and in what context. Which is yet another argument for the relevance of considering the broader socio-technical context. However, there seem to be central elements to the idea of security, which were highlighted when talking about the term ‘security algorithms’ and what their purpose is. Firstly, security algorithms are used to assess risk and predict crime, and secondly, they are tasked with keeping society safe. Putting these elements together, the value of security means: being safe from risks arising from potential crime. Translated to the context of predictive policing, security is defined as: *the extent to which predictive policing practice keeps people safe from risks arising from potential crime.*

2.3.2 Main Concerns regarding Predictive Policing

Most of the major concerns with regards to predictive policing deal with the topic of discrimination. On the topic of how predictive policing becomes discriminatory, Bayrne (2017), gives an account of how the practice of predictive policing could exacerbate social inequalities. In her paper she talks about three ways in which predictive policing can cause the reproduction of inequalities. Firstly, surveillance on individuals with criminal records gets deepened. As predictive policing is based on previous crime data, people who have been in conflict with law enforcement more, as a result of racial bias, may find themselves under even more scrutiny. Which is now justified on the basis of the supposed ‘objectivity’ of algorithms – this is an algorithmic form of confirmation bias. This problem also occurs for place based predictive policing, as the use of historical data may cause the police to unequally target certain neighbourhoods, thus exacerbating neighbourhood inequalities.

Secondly, using predictive policing broadens the scope for the number of people the police can legitimately track. In the current situation, the police needs have a proven suspicion as to why someone may pose a danger and needs to be surveyed. Predictive policing can provide such suspicion based on historical data, which automatically means that you can be under suspicion not only because you have a criminal record but also

because you portray certain characteristics or even because you frequent a location where there is a high probability of a crime occurring. Assuming that certain minorities are more likely to be associated with characteristics or places that are highlighted in historical crime data, it could be that they are disproportionately targeted by a predictive policing algorithm (Bayrne, 2017). Again the use of biased historical data is problematic here as existing inequalities and discriminatory practices already existing in the data will be replicated in the predictions.

Lastly, the use of these technologies can cause people to avoid institutions where they can leave a digital footprint. Especially if the police uses data from other institutions such as financial or medical data. People could start avoiding these institutions out of fear for what this could mean for their digital record. This effect is also negatively biased towards social minorities as they are historically more likely to have been in contact with the criminal justice system (Bayrne, 2017). This discussion by Bayrne (2017) clearly shows that the use of historical data that already has discriminatory bias embedded in it, is one of the main reasons why predictive policing could exacerbate social inequality rather than alleviate it.

The problems described above are not unique to predictive policing but are actually a general problem related to algorithm based decision making. There are 5 ways in which algorithms can unintentionally be made discriminatory. These are each related to one of the following problem areas: “(i) how the "target variable" and the "class labels" are defined; (ii) labelling the training data; (iii) collecting the training data; (iv) feature selection; and (v) proxies (Zuiderveen Borgesius, 2018, p.10).” In terms of the first problem, target variables define what an algorithm is looking for and class labels “divide all possible values of the target variable into mutually exclusive categories (Barocas and Selbst, 2016, p.678).” These can become discriminatory when these target variables and class labels are unequally affecting certain protected classes such as minorities (Barocas and Selbst, 2016; Zuiderveen Borgesius, 2018). In terms of predictive policing, an example could be when the police looks for the target variable ‘likely to commit a crime’ and uses the class label ‘lower income’. If a certain minority has on average a lower income and is therefore disproportionally targeted, you have a discriminatory model.

The second and third problems are both concerned with training data and are related to the examples by Bayrne (2017). Training data can be discriminatory due to the data being based on discriminatory human decisions. Putting this in a predictive policing example, if the training data consists of previous arrest reports and police officers have

been disproportionately arresting certain minorities, then the predictive algorithms will reproduce this bias and most likely disproportionately target these same minorities. The other way in which training data could be made discriminatory is in the data collection process. Specifically when data collected falsely over or under represent a certain social group. To give an example, if minorities are less likely to make reports of domestic violence then it could be that they are underrepresented in the training data, resulting in a situation in which minorities are less likely to be labelled as ‘at risk of domestic violence’ even though they might be more at risk (Barocas and Selbst, 2016; Zuiderveen Borgesius, 2018).

The fourth problem is related to the class labels that are selected as part of the analysis. If an algorithm is to make a prediction it cannot take into account all the possible variables. As such, a predictive model has to be created which is always a simplification of the real world. The creators of this model will have to make choices as to what features to include in the analysis. If this selection of features is in any way discriminatory then the results will also be (Barocas and Selbst 2016; Zuiderveen Borgesius, 2018). A simple example would be if the police would select ‘has a migration background’ as a class label. A more nuanced example could be if the police would select ‘level of education’ as a class label, targeting specifically people with a lower education. If minorities are overrepresented in lower education, than the model can turn out discriminatory predictions.

The fifth and final problem is proxies. This refers to a situation in which a seemingly objective characteristic, such as where someone lives or where someone went to school, becomes a proxy for a discriminatory characteristic (Barocas and Selbst, 2016; Zuiderveen Borgesius, 2018). If it is determined that ‘being from a certain postal code’ is considered as a characteristic of ‘at risk of committing a crime’, and a certain social minority is much more likely to live at that postal code, then that minority could become disproportionately targeted and the ‘postal code’ can become a proxy for the characteristic ‘social group’.

All of these instances in which an algorithm is made discriminatory unintentionally seem to be linked to the process of parametrisation’ which was mentioned back in the section on predictive policing. Remember that this refers to “the process of choosing independent variables that represent relevant aspects of the subject-matter problem (Hälterlein, 2021, p.3). ” The examples shown above, illustrate that there are many ways in which the human decisions involved in parametrisation can actually lead

to a discriminatory picture of crime being painted. But also that bias does not only enter the algorithm at the point of choosing the parameters but might already be present in historical data or established practices. This once again confirms the importance of considering the larger socio-technical context when trying to understand how broad explainability as a practice could interact with different values.

Most of the time, this discriminatory bias ends up in the algorithm unintentionally, but if it is done intentionally, it constitutes a sixth way in which algorithms can be made discriminatory (Zuiderveen Borgesius, 2018). Because there are so many ways in which predictive policing can become unintentionally discriminatory, there are many authors who emphasize the importance of values such as accountability and transparency in order to diminish the number of discriminatory practices (Bennet Moses & Chan, 2018; Samek & Müller, 2019; Meijer & Grimmelikhuijsen, 2020; Ferguson, 2017). Guidelines on AI ethics, often concerned with preventing discrimination and establishing a fair use of algorithms, also mention accountability and transparency (Hagendorf, 2019).

If we want to define the value of accountability, we can look at Bennet Moses and Chan (2018), who give a good explanation of the nature of accountability. Referring to work by Bovens et al. (2014), they explain that there are three important aspects to accountability. Namely that accountability involves: “(1) the provision of answers (2) to others with a legitimate claim to demand an account, (3) with consequences (p.817).” This is also what Wieringa (2020), who conducted a systemic literature review on algorithmic accountability, understands when it comes to accountability. Therefore, it can be concluded that accountability in the context of predictive policing means: *the extent to which the use of predictive policing algorithms can be assessed and consequences imposed based on those assessments.*

When discussing the need for more accountability, some authors also voice concerns regarding responsibility. Specifically, they are concerned with how responsibility is divided between the system and those who develop and use it (Bennet Moses and Chan, 2018). According to Meijer and Grimmelikhuijsen (2020), responsibility has five key elements: “(1) ethical judgment, (2) based on values, (3) and perceptions of relevant facts, (4) to enact a duty of care (5) through responsive pathways (pp.9-10).” They explain that responsibility is concerned with the duty of people to behave according to ethically respectable values, which also implies that people can be assessed based on whether they have acted in such a way. Responsibility in the context

of predictive policing can therefore be defined as: *the extent to which the duty of care for the proper use of the predictive policing algorithm has been clearly allocated.*

Transparency, can be understood as “the availability of information about an actor allowing other actors to monitor the workings and performance of this actor (Meijer, 2014, p.511).” Transparency, as such, is not a complex concept and, in the context of predictive policing, can be described as: *the extent to which information with regards to the whole predictive policing practice is made available.* However, as has been addressed briefly in the previous section, there is a discussion in academic literature on whether predictive policing increases or decreases transparency. Some proponents of predictive policing claim that the transparency of police departments using predictive policing is better when compared to those who do not. This would be, because the data which underlies the predictions that inform their decisions is readily available for scrutiny (Ferguson, 2017). On the other hand, there are also authors who doubt whether the availability of this data really does increase transparency. They specifically doubt the comprehensibility of the data for possible scrutinizers (Jansen & Van den Hoven, 2015).

Another important value is revealed here which is comprehensibility. Jansen & Van den Hoven (2015) argue that the complexity of machine-learning algorithms is so high that it becomes very difficult to understand the logic that is used in decision making, and the biases that are inherent in this logic. It can therefore be said that, in the context of predictive policing, the problems surrounding transparency are not necessarily only related to a lack of information but also to a lack of comprehension. Meijer (2014), also follows this line of reasoning, arguing that transparency, insofar as it entails an increase in available information, is only effective in facilitating accountability if there are people capable of interpreting and assessing the information. Furthermore, remember that, in order to fully determine whether a decision was made properly, not only the decision itself should be under scrutiny, or the data on the basis of which the decision was made, but the entire process – from data collection and storage to its use in the algorithm and the eventual decision output and action (Bennet Moses & Chan, 2018). Comprehensibility can therefore be understood as – *the extent to which information about predictive policing practice can be understood by the recipient.*

Explainability is also often named as an important value which combats discrimination caused by algorithmic decision making (Gilpin et. al, 2019; Samek & Müller, 2019; Weller, 2019). Alikhademi et al. (2021) for example argued that explainability is needed to provide more insight into the decision making process. Sandhu

and Fussey (2020) argue that inhibiting the explainability of predictive policing technologies also inhibit their accountability. Finally, Haque, Weathington, Chudzik and Guha (2020) recognize that a lack of explainability could negatively impact peoples trust in predictive crime mapping. Explainability however, will be thoroughly discussed later in this research background and will therefore not be defined yet.

Two final values are often named as important with regards to discrimination in algorithmic decision making: trust and fairness (Samek & Müller, 2019; Weller, 2019; Hagendorff, 2019; Meijer & Grimmelikhuijsen, 2020). Trust can be seen as an important condition for government legitimacy (Meijer and Grimmelikhuijsen, 2020), and thus the legitimacy of predictive policing. However, trust is not a necessary condition for citizen acceptance or obedience. What trust can do is make it easier for governments to get citizens to comply with their policies and accept their decisions. To define trust, “trust should be understood as a multidimensional concept that consists of citizens perceptions of government competence, benevolence and integrity (Meijer and Grimmelikhuijsen, 2020, p.6).” Trust, in the context of predictive policing, as such, can be understood as: *the extent to which people believe the predictive policing algorithm is treating them fairly and is working for their benefit.*

Fairness seems to be most directly linked to people’s concerns with regards to discrimination in algorithmic decision making generally (Rai, 2019; Weller 2019), and predictive policing specifically (Veale, van Kleek, & Binns, 2018; Lepri et al., 2018). This is because fairness is mostly understood as the absence of discrimination or bias in algorithmic decisions (Corbett-Davies et al., 2017; Veale, van Kleek, & Binns, 2018; Lepri et al., 2018; Rai, 2019; Weller 2019). Fairness, in the context of predictive policing can therefore be understood as – *the extent to which an predictive policing algorithm is considers everyone on the same basis.* What is interesting to see from the definition of fairness and trust is that, in contrast to the other values that were named in this section, these are not seen as important values when fighting discriminatory practices in predictive policing. Rather, they seem to be more indicative of whether people judge these practices to be discriminatory or not.

A final important concern that people have with regards to predictive policing deals with possible infringements on privacy. Privacy is identified as a value in many different guidelines on AI ethics (Hagendorff, 2019). Privacy and predictive policing have a complicated relationship because the former is concerned with the protection of personal data and the latter inherently uses personal data in order to form predictions. In

predictive policing practice this causes certain problems to emerge: (1) the right of people to know about crime and the victims right to privacy, (2) researchers are interested in the data used by the predictive algorithms but some of this data is personal and should therefore remain private, (3) unexpected negative social outcomes resulting from the sharing of crime data, and (4) intentional sharing of crime data might make police departments vulnerable for data leaks (Wartell & McEwen, 2001). As such, the problem of privacy in the context of predictive policing deals with whether private data is secured and cannot be traced back to individuals. Privacy, therefore, can be understood as: *the extent to which private data used for predictive policing is secure and untraceable*.

At this point 12 important values have been identified in the context of predictive policing. Furthermore, 11 of them have been defined, only leaving explainability without a distinct definition. This value will be discussed and defined in the next section. This section will also discuss the current understanding of how explainability could foster values, why there is a need to specifically talk about broad explainability, and how this practice might interact with the values identified in this section.

2.4 Explainability: A Broader Perspective

Although the concept of explainability is relatively new, research on the interpretability of intelligent systems has a history of more than 50 years (Hansen & Rieger, 2019). The concepts of interpretability and explainability have often been used interchangeably in literature, but an important distinction has been made in recent years. Interpretability can be understood as a useful starting point for explainability, pertaining to the practice of comprehending what a system exactly did to produce an output (Gilpin et al., 2019). Nevertheless it is insufficient as a concept to solve the problems associated with black box systems as it only covers finding out how a system came to an output, regardless of whether this can be communicated to others. To illustrate, producing a flowchart of all the different logical steps that an algorithm took to produce an outcome can hardly be regarded as understandable from a practical point of view, though it would make an algorithm interpretable.

Explainability, as such, goes further than interpretability as it is also concerned with whether interpretations can be communicated to others in a comprehensible way (Gilpin et al., 2019; Samek & Müller, 2019). In this sense, the concept of explainability is much broader, covering not only how a system arrives at different outputs but also whether this can be communicated in a meaningful way. The benefit of explainability

over interpretability is therefore that it necessitates meaningful comprehension on the part of the people who receive an explanation.

Explainability, rather than interpretability became much more important due to the increased complexity of algorithm based systems. Systems, such as those based on machine learning algorithms, deep learning algorithms and complex neural networks, grew increasingly more complex to the point where these systems became black boxes. Sometimes even to those who had to work with them directly (Rai, 2019; Bertossi & Geerts, 2020; Samek & Müller, 2019). It was for this reason that explainability as a value and so called Explainable AI (XAI) methods (i.e. methods of realizing explainability) attracted more attention from scholars and practitioners, as these could supposedly provide the transparency, justification and traceability necessary to foster trust in such black-box systems (Shaban-Nejad, Michalowski, & Buckeridge, 2021). Indeed, some authors argue that explainability has a direct relationship with values such as trust, transparency and fairness (Hansen & Rieger, 2019; Shaban-Nejad, Michalowski, & Buckeridge, 2021; Samek & Müller, 2019; Weller, 2019). The ultimate benefit of explainability therefore, is its potential to foster certain values by improving comprehension.

Besides comprehension, two other elements are central to our current understanding of how explainability could foster values. This is best showcased by considering the explanation of explainability that Meijer and Grimmelikhuijsen (2020) give in their paper on algorithmization. According to them, explainability “concerns the substantive reasons for a decision: on what grounds was the decision taken and how does this relate to legislation and other formal rules and policies (p.13). Other discussions on explainability almost always include these three elements: comprehension, explaining the reasoning behind an algorithmic decision, and enabling scrutiny.

For example, Samek and Müller (2019), explain that explainability can help make the decision making of an algorithm more understandable, breaking open the black box of algorithmic decision making, and providing a basis for verifiability. Similarly, Rai (2019), argues that ‘Explainable AI’ (XAI) is “the class of systems that provide visibility into how an AI system makes decisions and predictions and executes its actions. XAI explains the rationale for the decision-making process, surfaces the strengths and weaknesses of the process, and provides a sense of how the system will behave in the future (p.138).” As can be seen, each of these explanations include references to comprehension, explaining the reasons behind algorithmic decisions, and scrutiny. The

main value of explainability, however, seems to lie in its assumed relationship to comprehension. By revealing the reasons behind an algorithmic decision and thereby improving peoples comprehension of how these decisions are made, these systems can be scrutinized. As such, explainability as a value can be understood as: *the extent to which information can be communicated in a way that can be understood by recipients.*

When it comes to methods of realizing explainability, there are only so called XAI methods to choose from. Rai (2019) provides an overview of different forms of XAI explanations and distinguishes between global and local explanations. A global explanation aims to explain the entire system, or rather the model underlying the system, whereas a local explanation aims to explain only a single output of a system. On top of this, Rai (2019) also distinguishes between model-specific and model-agnostic explanations. Model-specific explanations are aimed at incorporating “interpretability constraints within the inherent structure and learning mechanism underlying deep learning models (p. 138).” On the other hand, model-agnostic explanations are aimed at providing explanations based only on the input and output of a black-box model. With these distinctions in mind, Rai (2019) arrives at four different types of explanations: (1) model-specific global; (2) model-specific local; (3) model-agnostic global; (4) model-agnostic local. Each of these types of explanation covers its own set of technical methods that can provide explanations that fit the type description, e.g. interpretability constraints, attention mechanisms, diagnostic techniques and the Local Interpretable Model-Agnostic Explanation (LIME) technique.

This discussion of XAI methods certainly explains the different ways in which a system can be made explainable, but they do not contain hints with regards to how these explanations could foster different values. With regards to this, Rai (2019) only explains that explanations are important for fostering trust and they assume that making the system explainable will allow for better explanations and thus foster more trust.

Other authors who discuss methods of realizing explainability also provide no actual explanation on how explainability could foster certain values. The methods of interpreting AI systems, outlined across several different articles in the edited book on Explainable AI by Samek, Montavon, Vedaldi, Hansen and Müller (2019), are good examples of this trend. Samek and Müller (2019) are the best example of this, they argue that providing explanations, and thereby integrating people into the decision-making process, will foster trust. Even if these explanations do not provide additional information or are fully comprehensible for the receiver, the fact that people receive an explanation

would still foster acceptance and provide a basis for informed consent. They however, do not provide an account of how this would actually work or whether this is always the case. We also see this is happen with authors who specifically argue for the importance of explainability in the context of predictive policing. Sandhu and Fussey (2020), as well as, Haque, Weathington, Chudzik and Guha (2020), make arguments in favour of explainability based on the need for more fairness and trust but they do not explain how explainability would foster these values in practice. It seems to be a trend in literature on explainability to assume that explainability will allow for better explanations and that this will foster certain values.

Looking at the methods of realizing explainability described above however, there are some serious doubts as to whether these technology-centred explanations could really foster values. Especially when viewing these explanations from the holistic perspective on predictive policing practice and considering the wider socio-technical context. From this point of view, the explanations outlined above only cover the technology which should be considered as only one part of a larger crime fighting strategy. Furthermore, these explanations would only cover the substantive reasons of the system for giving a certain prediction. But it would not give the substantive reasons for the actions which were taken on the basis of the prediction, or the substantive reasons behind the selection of the data that is used. Both of which could be more relevant when fostering e.g. trust. Lastly, it is questionable whether a highly complex technology-focussed explanation like described above, could even be understood by people who do not have a background in computer science, let alone foster a sense of fairness or trust.

This is also what Miller (2019) argued in his paper on the explainability of AI. In his paper, he argued that the developers of AI who are producing the various methods of making algorithms explainable are unsuited to judge how useful such explanations are for lay people. Calling it ‘the inmates running the asylum’. For these reasons, he argued that the field of explainable AI should take lessons from social research focussed on how humans explain things to other humans. According to him, this is necessary if the goal is to design intelligent systems that are able to explain how they work to average people. It can be concluded therefore that explanations covering only the algorithm and how it produces decisions might be unsuitable for fostering any kind of value due to the fact that they might not be comprehensible or cover irrelevant information.

To summarize, the literature on realizing explainability has been very focussed on how to explain the substantive reasons behind algorithmic decisions, and just assumes

that the other two elements of explainability – improving comprehension and enabling scrutiny – will automatically follow in practice and that therefore explanations will foster values. Hagendorff (2020) also concluded that explainability is mainly implemented in a mathematical way in the form of technological solution. Considering these assumptions from the larger socio-technical context however, it becomes clear that this cannot be assumed and that explanations that only explain the technology might be irrelevant or incomprehensible. As such, our current understanding of explainability as a practice, which is centred on explaining the technology only, is unsuitable as a basis of explaining how explanations interact with different values and how these interactions might foster these values. As such, this thesis argues that it is necessary to broaden the concept of explainability and focus on explaining not only the technology, but also other aspects of the wider socio-technical context as well. In line with this, this thesis proposes a new concept: broad explainability.

As such, this thesis will distinguish between two different approaches to practicing explainability: (1) Technology-centred explainability – *the act of giving explanations aimed at explaining the technology and how it makes decisions*, and (2) Broad Explainability – *the act of giving explanations that are aimed at explaining multiple aspects of the socio-technical context*. Please note that broad explainability, defined as such, can still include an explanation of the technology and how it functions. Of these two, broad explainability seems to be most promising with regards to fostering values in a practical context, as it is not limited like technology-centred explainability and could therefore cover more relevant information and, as such, has a better chance at actually improving comprehension and allowing for meaningful scrutiny. This is why this thesis focusses specifically on broad explainability and how this practice could foster different values.

At this point, it must be stressed that, despite the fact that broad explainability might convey more relevant information than technology-centred explainability, this still does not guarantee that these explanations actually foster certain values because they do not guarantee an improvement in comprehension. Furthermore, even if it is assumed that explanations foster better comprehension, it is still not guaranteed that they have a positive impact on values. Consider for example Weller (2019), who explains that there are scenarios in which increased transparency can actually cause discriminatory behaviour, showing that sharing certain information can actually endanger certain values, in this case fairness. The act of giving explanations inherently involves the sharing of

information, and especially if we focus on broad explainability, these explanations will hypothetically contain a lot of information pertaining to the entire socio-technical context. As such, it must be concluded that the relationship between broad explainability and other values is not strictly fostering, but that giving explanations might have to be balanced against realizing certain values. In which case we can say that the relationship between broad explainability and a value is balancing.

Two potential ways in which broad explainability could relate to these values can therefore be outlined, namely: (1) a fostering relationship, meaning that giving broad explanations could potentially foster these values, and (2) a balancing relationship, meaning that giving broad explanations might endanger certain values and must therefore be balanced against this value. As such, when answering the research question, this thesis will discuss whether the relationship between broad explainability and a particular value can be fostering and at what point it becomes more balancing.

Furthermore, one last observation has to be added. If we consider the larger socio-technical context of predictive policing, we cannot assume that these values exist in a vacuum. As was hinted at, comprehensibility seems to have links to transparency and definitely relates to explainability. Similarly, responsibility and accountability, as well as trust and fairness, also seem to be closely related. As such, this thesis will also look at which other values are related to a particular value, in the context of fostering that particular value by means of broad explainability.

2.5 Fostering Values with Broad Explanations

In the last section, explainability as a value was defined as *the extent to which information can be communicated in a way that can be understood by recipients*. With that, all 12 values that were identified as important in the context of predictive policing have been defined and can be summarized in a table:

Value	Definition
<u>Explainability</u>	The extent to which information can be communicated in a way that can be understood by recipients.
<u>Efficiency</u>	The extent to which police resources are optimally allocated in spatial and temporal sense.
<u>Effectiveness</u>	The extent to which police operations are successfully reducing crime rates.
<u>Accuracy</u>	The extent to which a predictive algorithm makes correct predictions

<u>Security</u>	The extent to which predictive policing practice keeps people safe from risks arising from potential crime.
<u>Accountability</u>	The extent to which the use of predictive policing algorithms can be assessed and consequences imposed based on those assessments.
<u>Responsibility</u>	The extent to which the duty of care for the proper use of the predictive policing algorithm has been clearly allocated.
<u>Transparency</u>	The extent to which information with regards to the whole predictive policing practice is made available.
<u>Comprehensibility</u>	The extent to which information about predictive policing practice can be understood by the recipient.
<u>Trust</u>	The extent to which people believe the predictive policing algorithm is treating them without prejudice and is working for their benefit.
<u>Fairness</u>	The extent to which a predictive policing algorithm considers everyone on the same basis.
<u>Privacy</u>	The extent to which private data used for predictive policing is secure and untraceable.

Last section also differentiated technology-centred explainability from broad explainability and explained that this thesis focusses specifically on broad explainability – which refers to the act of providing explanations covering the entire socio-technical context. This however, leads to an awkward situation in which explainability has both been identified as an important value in the context of predictive policing, and as a practice that might foster this value. It is therefore important to explain that this thesis identifies these two things – explainability as a value and broad explainability – as two distinct things that can be impacted by one another, i.e. broad explainability might foster explainability as a value or need to be balanced against it.

The last section ended with the identification of two potential ways in which broad explainability could relate to the values described above (fostering & balancing), which will form the basis of answering the first part of the research question. However, as the research question shows, this thesis also aims to study how practicing broad explainability could actually foster these values. Therefore, in these final parts of the research background, the researcher proposes a framework with which to study how broad explainability could actually be practiced in order to foster the values identified previously. The framework will be inspired by work from: Weller (2019), Samek and Müller (2019), and Meijer and Grimmelikhuijsen (2020). In more practical terms this means that the next section will discuss what factors to look at in order to determine what the right explanation is for fostering a certain value in the context of predictive policing. This will also include a discussion of which aspects of predictive policing, besides the

technology, could be explained, if we consider the larger socio-technical context of predictive policing from a holistic point of view.

2.5.1 Determining the Content of Explanations

To determine what the right explanation is for fostering different values, some lessons can be learned by considering different types of transparency. In his paper, Weller (2019), lists different types of transparency which contrast each other on goal, intended audience and likely beneficiary:

#	Audience	Beneficiary	Goal
1	Developers	Society	To understand how their system is working, aiming to debug or improve it: to see what is working well or badly, and get a sense for why.
2	Users	Society	To provide a sense for what the system is doing and why, to enable prediction of what it might do in unforeseen circumstances and build a sense of trust in the technology.
3	Society	Society	To understand and become comfortable with the strengths and limitations of the system, overcoming a reasonable fear of the unknown.
4	Users	Society	To understand why one particular prediction or decision was reached, to allow a check that the system worked appropriately and to enable meaningful challenge (e.g. credit approval or criminal sentencing).
5	(Legal) Experts	Society	To provide the ability to audit a prediction or decision trail in detail, particularly if something goes wrong (e.g. a crash by an autonomous car). This may require storing key data streams and tracing through each logical step, and will facilitate assignment of accountability and legal liability.
6	(Safety)Monitors	Society	To facilitate monitoring and testing for safety standards.
7	Users	Deployer	To make a user (the audience) feel comfortable with a prediction or decision so that they keep using the system.
8	Users	Deployer	To lead a user (the audience) into some action or behavior – e.g. Amazon might recommend a product, providing an explanation in order that you will then click through to make a purchase.

Adapted from “Transparency: Motivations and Challenges,” by A. Weller, in W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (pp. 23–40). Springer International Publishing.”

As can be seen from this table, type 1-6 are of general benefit to society, but, Weller (2019) adds, only on the condition that the explanations are given faithfully, meaning given accurately and without omission of important details (Weller, 2019). Type 7 and 8 are of a much more manipulative nature and are only beneficial for those who

deploy a certain system. What Weller (2019) wants to illustrate by outlining these different types of transparency is that, based on parameters such as audience, beneficiary and goal, transparency can take many forms and as such, it is important to keep these things in mind when looking for the optimal type of transparency.

Discussing specifically the content of explanations, Samek and Müller (2019), identify almost the same distinguishing factors as Weller (2019). They explain that explanations can differ based on the recipient, the information content and the intended goal. Although they focus only on technology centred explanations, they show that different recipients may need different levels of detail in their explanations as well as different aspects of the technology explained. This links to the information content, on which these authors argue that different explanations, depending on the intent, might need to focus on different aspects in order to be advantageous. Finally, this in turn, links to the goal of an explanations. On which these authors argue that the goal of an explanation can differ based on (1) how one intends to provide an explanation, and (2) what one wants to use the explanation for.

Now as has been said, these authors argue specifically about different types of explanations with a focus on explaining the technology. Nevertheless, when considering these arguments in the holistic view on predictive policing, it seems that these claims hold up. First of all, different people could ask for an explanation of the predictive policing process, e.g. citizens, police officers, and the judiciary. As such, it seems that we have to take into account who we are giving an explanation to. Secondly, as has been shown in, there are different aspects to the use of an algorithms, and as such, there are many different aspects to the predictive policing process which could be explained. It is therefore also important to take into account what the content of explanations could and should be. Finally, on the goal of explanations, this thesis is looking at how explanations could potentially foster different values in the context of predictive policing. Although this seems like a singular goal, this can actually take many forms, depending (1) on which value you want to foster and (2) from whose perspective you want to foster it. Fostering transparency from the perspective of citizens means something different than from the perspective of police officers. Furthermore, each of these factors are interdependent, as was shown in the previous paragraph. It should be clear therefore, that, when determining what the right explanation could be, these three factors – recipient, content, and goal – have to be taken into account as interdependent variables.

Now Weller (2019) identified another factor: likely beneficiary. This factor will however not be included in the scope of this thesis. As can be seen from the 8 examples by Weller (2019) in table 2.2, the potential beneficiary can often be described in general terms like ‘society’ and, this thesis argues, that this factor is less important for determining the specific type explanation than the intended audience and the intended goal are. This is because the beneficiary of an explanation always has to be guessed beforehand and can only be confirmed after an explanation has been given. As such, when attempting to determine what the right type of explanation is, focusing on the likely beneficiary does not seem to be productive. Narrowing the number of variables considered in this thesis down further, it must also be said that the goal of the explanations, considered in this thesis, is also a given. This thesis looks at how explanations could foster values in the context of predictive policing, which automatically provides the intended goal of the explanations – fostering the different values that were identified as important in the context of predictive policing.

This leaves us with two factors – from now on referred to as ‘factors of explanation’ – to consider when determining what kind of broad explanations could foster the different important values in the context of predictive policing. These are (1) the intended audience and (2) the right content. One more factor has to be added to these two for the purpose of this thesis. As was argued for in the previous section, we need to take into account the other related values when studying how broad explainability fosters a certain value. As such, this thesis will also look at a third factor of explanation which is (3) values that are related to a particular value, in the context of fostering that particular value by means of broad explainability.

This leaves one question open, however: what could the possible content of explanations be – taking into consideration the entire socio-technical context from the holistic point of view of predictive policing. As was shown before when discussing the factor of content. Content is variable because there are different aspect to the use of an algorithm that can be explained. Additionally, it must be noted that no explanation could possibly contain all different aspects to the use of predictive policing, and therefore, a selection of relevant aspects, with regards to the goal of an explanation, should be made. The next section will therefore outline what different aspects of the use of predictive policing could be identified along with the content of explanations could be structured.

2.5.2 Algorithmization: Aspects of the Socio-Technical Context

In order to determine what different aspects of the socio-technical context of predictive policing could be explained from a holistic point of view, it is helpful to look at the concept of algorithmization. Algorithmization is a concept coined by Meijer and Grimmelikhuijsen (2020) and refers to the process of “organizational change around the introduction of algorithms (p.1)”. Their research is part of the larger trend in scholarly literature which seeks to move emphasis away from questions surrounding efficiency and effectiveness and put more focus onto questions and challenges surrounding fairness, discrimination, privacy and generally avoiding unintended negative consequences of the use of algorithms (Hoffman, 2019; O’Neil, 2016; Eubank, 2018; Gerards, 2019). As Meijer and Grimmelikhuijsen (2020) point out, the concerns of these scholars go further than the mere implementation of algorithms. Therefore, addressing a broader number of issues associated with organizational change and how this can enable the ethical use of algorithms. Because they broadened the scope of their research to focus, not only on the technology itself, but also on the surrounding organizational change, Meijer and Grimmelikhuijsen’s (2020) concept of algorithmization provides a perfect starting point for theorizing about the different aspects of the socio-technical context of predictive policing which could be included in an explanation fostering a certain value.

Meijer & Grimmelikhuijsen (2020) explain that algorithmization in the public sector consists of 6 components: (1) Technology – referring to the algorithm itself either as a standalone system or a system integrated into the organisational infrastructure, (2) Expertise – the level of expertise available in an organisation with regards to the use of algorithms, (3) Information Relations – the effects of the algorithm on the information relations in an organisation caused by its use of old information, production of new information and use of information from outside sources, (4) Organizational Structure – possible new departmental collaboration or organizational control structures resulting from the use of the algorithm, (5) Organizational Policy – policies surrounding the algorithm pertaining to e.g. transparency, responsibility and maintenance, and (6) Monitoring and Evaluation – methods of monitoring and evaluating foreseen and unforeseen consequences of the use of the algorithm. Each of these components could be part of an explanation from the perspective of broad explainability.

One thing however, seems to be left out of the components of algorithmization outlined by Meijer and Grimmelikhuijsen’s (2020). None of the components seem to

include the interplay between the outputs of an algorithm and human decisions which results in certain actions being taken. This makes sense as the concept of algorithmization was not designed to describe the actual use of the system but rather the context in which it is used. It does however, mean that the components of algorithmization do not describe a very important aspect of the socio-technical context from a practical point of view. In other words, as they are now, the components of algorithmization do not completely cover all relevant aspects of the socio-technical context of predictive policing when viewed from the holistic perspective that is taken in this thesis.

Algorithmization component 5 – Organisational Structure – approaches this idea the most, but does not quite cover it. In their own words, Meijer and Grimmelikhuijsen (2020) explain this component as follows:

The use of the algorithm will often result in new collaborations between different departments. The algorithm can also result in new forms of organizational control when implementation of processes is dictated by the algorithm. (p.7).

It can be seen that this component accounts for new processes and relations being created but does not explicitly cover the interplay between humans and the technology in the larger socio-technical system. It is for these reasons that a seventh component of algorithmization will be added for the purpose of this thesis, namely: (7) Socio Technical relations. Explained as: *the interplay between the outputs of an algorithm and human decision making with regards to these outputs which result in certain actions being taken.*

These 7 components of algorithmization will serve as a first list of the different aspects of predictive policing along which the content of explanations could be structured. However, it must be noted that they do not and cannot possibly cover the entirety of the socio-technical context of predictive policing and everything that is involved in this – as is called for when adopting a truly holistic view on predictive policing. For example, if one is to consider the development of the predictive policing tool as part of the predictive policing process then this aspect is not covered by these 7 components. But, in order to somewhat limit the scope of this thesis and keep the topic manageable from a practical point of view, this thesis will consider only these 7 components of algorithmization as possible content for explanations that could be given with regards to broad explainability.

3 Research Design and Methodology

3.1 Research Design

This thesis has followed a qualitative research design, which was chosen because of the novel point of view from which this thesis looks at predictive policing and explainability and the, therefore, lack of previous research. Also meaning that this thesis is highly explorative. As Patton (2015) explains, “qualitative inquiry is particularly oriented towards exploration, discovery, and inductive logic. Inductive analysis begins with specific observations and builds towards general patterns (p.122).” Based on the research question and literature review, it should be clear that this thesis intends to study expert perspectives on how broad explainability interacts with other values in the context of predictive policing, with the aim of drawing general conclusions with regards to whether these interactions are fostering or balancing and how different values could be fostered by giving broad explanations. A qualitative approach is therefore, best suited for this thesis, hence the choice for a qualitative research design.

Before definitively landing on the research question shown in the introduction and the methodology described below, a systemic literature review was conducted following the methods outlined by Webster and Watson (2002). This was necessary because this thesis has the aim to discuss and contribute to, the previous academic debate on explainability and its potentially fostering relationship to certain values. This systemic literature review, therefore, initially aimed to combine different streams of literature on: ‘Explainability’, ‘AI’, ‘Explainable AI’, and ‘Algorithmic Decision Making’ in the ‘Public Sector’. Based on these five main concepts, a database search was conducted. The databases that were considered were: Web of Science, Limo (Database of KU Leuven), and Google Scholar. These databases cover a broad range of journals and research fields and therefore, can provide a broad insight into past research on these topics from different perspectives. This is valuable when aiming to conduct a complete review as relevant literature on a topic is not limited to one specific field or journal (Webster & Watson, 2002). After this first search, 13 articles were selected as relevant after reading their abstract or introduction. Special attention was given to the recency of the articles with more recent articles being seen as more relevant.

After considering this first batch of articles, the specific context of predictive policing was chosen, which necessitated the consideration of some additional relevant concepts, which were: ‘Predictive Algorithms’, ‘Predictive Policing’, ‘Fairness’,

‘Discrimination’, ‘Accountability’, ‘Transparency’. This second search yielded an additional 21 articles. Following the analysis of this second batch of articles a backwards search was conducted in all the previously selected articles which yielded a another 16 articles. Finally, 4 articles were recommended and, upon inspection, deemed relevant and included. All 55 selected articles are depicted in the concept matrix in appendix A which was created based on Webster and Watson (2002).

As mentioned earlier, while conducting the (systemic) literature review, the research question and methodology were refined. Merriam and Tisdell (2016), explain that when designing a qualitative study, writing the literature review, forming the problem statement and developing the theoretical framework is an interactive process. After which a sample to study can be selected. As such, it clear that when designing a qualitative study, the literature review, research question and methodology are not created sequentially but simultaneously, with each element impacting the others.

3.2 Methodology

As Patton (2015) argues in their book on qualitative research, the goal of the research is the main driver behind the design of the research. With regards to this thesis, the research goal is summarised in the research question, from which it becomes clear that access to information pertaining to actual predictive policing practice is necessary in order to answer it. For this reason it might be surprising that this thesis follows an exploratory expert interview methodology, rather than an exploratory case study methodology. To answer this concern, it should be mentioned that an exploratory single case study methodology was considered and pursued. However, the switch to the current methodology was necessary for two reasons: (1) unsuitability of the selected case, and (2) lack of willingness to participate. One initial interview was conducted while pursuing this methodology, based on which the usability of the selected case came into question. Afterwards, during correspondence with several representatives of the specific case it was discovered that a number of initial assumptions with regards to the case were faulty, which completely disqualified the case. Finally, the representatives of the case also expressed no willingness to participate in the research.

In order to still be able to gain access to the necessary information regarding predictive policing practice, an exploratory expert interview methodology was chosen. As Bogner and Menz (2009) explain, the exploratory expert interview is a tool used in both quantitative and qualitative research projects and can help “establish an initial

orientation in a field that is either substantively new or poorly defined (p.46)”; which is the case in this thesis. Used like this, the expert is considered as a source of information on the subject that is actually being studied. They do so in the capacity of someone who possesses ‘contextual knowledge’ (Bogner & Menz, 2009). This means that these people earn the status of experts based on their knowledge of a given research field and/or their knowledge of structures, procedures and events in certain organizations (Littig, 2009). The shared context between experts ensures the comparability of the interviews (Meuser & Nagel, 2009). In this thesis the shared context between the experts is predictive policing.

The benefit of using an exploratory expert interview methodology, instead of an exploratory single case study methodology (which could also be combined with expert interviews), is that, by removing the specific context of the case and using predictive policing in general as the shared context, this thesis will be able to generate much broader insights with regards to its subject of study: broad explainability and how it could foster values. This is in a large part due to the fact that a broad range of different experts, with expertise spanning across different academic fields and different predictive policing projects, can be included in the research. On the downside, the insights that can be induced are much more shallow and largely hypothetical as it is not possible to go into much depth on specific practical examples during the interviews. This limitation and other limitations such as, but not limited to, those stemming from the chosen methodology and the specific set-up of the interviews, will be discussed at the end of this thesis.

3.2.1 Exploratory Expert Interviews

According to Bogner and Menz (2009), exploratory expert interviews should be conducted as openly as possible. However, they also emphasizes that it is advisable to structure the central dimensions of the interview along a topic guide. In this regard, the exploratory expert interview, as explained by Bogner and Menz (2009), is very similar to what is commonly known as a semi-structured interview. Merriam and Tisdell (2016), in their book on qualitative research, explain that most interviews in qualitative research are semi-structured interviews, and that this type of interview forms the middle ground between structured/standardized interviews and unstructured/informal interviews. Being the middle ground between these two extremes means that the questions for a semi-structured interview are, to some extent, predetermined. However, they are flexibly worded or they are a variety of more and less structured questions (Merriam & Tisdell,

2016). In practice, this means that before the interview, an interview guide/schedule is created which loosely outlines the topics that are to be discussed and flexibly words the predetermined questions. This provides the interviewer with the ability to gather similar information from all interviewees but also allows for enough flexibility to respond to the narrative constructed by the interviewees and explore new ideas (Merriam & Tisdell, 2016). The interview schedule for the interviews conducted in the context of this thesis can be viewed in appendix B.

As was mentioned, in order to ensure the comparability of the interviews, a shared context must be established between the experts (Meuser & Nagel, 2009). For this thesis, this context is predictive policing, however the view that this thesis takes on predictive policing is very specific and therefore it was deemed necessary to prepare a document with background information in which this thesis's holistic view on predictive policing is explained. Furthermore, this thesis aims to study the practice of broad explainability in the context of predictive policing. As the notion of broad explainability was developed in this thesis, it was also deemed necessary to include an explanation of this concept in the document. Finally, this document also contained a description of the specific topics that were to be discussed in the interview, consisting of two tables – one with the identified values and their definitions, and one with the components of algorithmization. This document was shared with the interviewees beforehand and served a dual purpose: (1) it would ensure a common context and common understanding of the main phenomenon's studied – broad explainability and predictive policing, and (2) it would minimize the need for explanations during the interview. This last point would also help ensure that all the necessary topics could be covered in the limited time of the interviews. Special care was taken to ensure that the document was merely informative and would not create biases with the experts before the interviews. It can be viewed in appendix C.

As for the specific set up of the interview, the interviews were roughly divided into two parts. The first part of the interview was aimed at discussing the relationship between broad explainability as a practice and the various values that were identified in the context of predictive policing. Specifically, it was discussed whether this relationship is more fostering or balancing from the perspective of the expert. The second part of the interview was concerned with discussing what kind of explanations could foster different values, looking at the three factors of explanation that were identified in the literature review. Because of time constraints, the list of values that would be discussed in the second part had to be limited and for that reason, the interviewees were asked to choose

the 5 values that are, according to them, most important to foster in the context of predictive policing. Subsequently, for each of the five values, it was discussed who the intended audience should be, which other values come are related when fostering a particular value, and which components of algorithmization should be covered in the explanation. With explicit consent of the interviewees, all interviews were recorded to allow for transcription at a later point.

During the interviews, which were all conducted via Microsoft Teams, the interviewer showed a document on screen which was filled in together with the expert. This document can be viewed in appendix D, and served the purpose of guiding the interview, as well as displaying some relevant information to aid the experts when coming up with their responses. Specifically, a table with the identified values and their definitions, along with a table containing the components of algorithmization and their descriptions were displayed. This information was also part of the background information that was shared with the experts beforehand. Besides this, the document contained dedicated tables in which the general answer of the expert per value discussed could be documented. Documenting the answers of the interviewees in this way constitutes good methodological practice, as it provides backup notes in case the recording equipment fails (Patton, 2015). Additionally, it also helped the interviewer to manage time and make sure that all the relevant topics were covered.

To maintain the semi-structured nature of the interviews, the researcher made sure to ask experts to elaborate on their answers and to ask unscripted follow up questions when deemed relevant. Furthermore, during the interviews the experts were left free to wander between the different values and into different topics. Though only to a certain extent as the researcher had to keep a tight schedule.

A final point to be mentioned is that, before definitely landing on the interview set up described above, two trial interviews were conducted. These interviews were conducted for two reasons. Firstly, to identify the best way of structuring the interviews so that all the relevant topics could be covered with enough depth, while also allowing the experts to freely discuss related topics and to ask follow up questions. This would also help ensure compatibility between the interviews. Secondly, these interviews helped to further define the criteria for the sample of experts that this thesis would study – this will be further explained in the next section. An overview of all the interviews conducted for this thesis can be found in appendix E.

3.2.2 Sampling Method and Expert Sample

The expert sample studied in this thesis was selected based on a combination of different nonprobability sampling methods: convenience sampling and snowball sampling. Additionally, a Google search was conducted based on predetermined criteria. Merriam and Tisdell (2016) explain that most qualitative research uses nonprobability sampling as the main method of selecting an appropriate sample. The most common form of nonprobability sampling is purposeful sampling, which means that a sample is selected based on the purpose of the study. In more practical terms, this means that the researcher determines a set of criteria which includes the attributes that a potential candidates must have in order to be included in a sample (Merriam & Tisdell, 2016).

On the two main sampling technique's used in this thesis, as the name describes, convenience sampling means that a sample is selected based on convenience. For example in terms of time, willingness and/or availability. Merriam and Tisdell (2016) explain that convenience sampling almost always plays a role during sample selection, but that it is important to have criteria besides convenience in order to establish credibility. Finally snowball sampling is a technique that looks for key candidates for interviews and asks these to provide other candidates that meet the predetermined criteria (Merriam and Tisdell, 2016).

These two methods of purposeful sampling are different from the other methods of purposeful sampling because they do not deal with the question of representativeness. To illustrate, over the years, scholars have started to differentiate between different kinds of purposeful sampling. A typical sample would be one that is representative of a certain group, otherwise one could select a sample based on atypical attributes. As a third option, a researcher might opt to go for the maximum variation in their sample, wanting to include a wide variety of instances of a phenomenon or group of people (Merriam & Tisdell, 2016).

Although these are much used purposeful sampling methods, none of these sampling techniques were deemed suitable for this thesis, specifically because they all deal with the question whether or not a sample should be representative of the pool of people from which a sample is selected. As Littig (2009) explains, sampling for expert interviews does not adhere to the conventional rules of representativeness in qualitative research. This is because experts are not part of a clearly defined group, as the status of expert is rewarded based on criteria determined by the research goals. Therefore, it cannot

be determined whether a selected sample is or is not representative of the entire pool of people who could have potentially been selected as an expert. Although this might seem like a limitation, it is not when one considers that the experts are not the subject of study, but sources of information pertaining to the actual subject being studied.

Turning now to the criteria that were used for selecting the sample. In the context of this thesis, these criteria will determine who will be counted as an expert and what kind of expertise they need to have in order to be included in the sample. As was explained before, in broad terms, an expert is someone who possesses contextual knowledge (Bogner & Menz, 2009). Meaning that they either have intimate knowledge of a subject from the perspective of a certain research field and/or they have knowledge of structures, procedures and events in certain organizations (Littig, 2009). Due to the complex and abstract approach that this thesis takes towards studying its topic, both of these types of knowledge were needed in the expert sample. As such, the decision was made to focus specifically on academics as experts. The first two criteria for selecting the sample were therefore: (1) a background in research on predictive policing, and (2) knowledge of predictive policing practice gained through this research or practical experience.

During the two trial interviews, in which two people were interviewed who had studied predictive policing from very different research fields – governance & innovation and criminal law & applied ethics – an interesting but obvious conclusion was drawn. Which was that looking from the perspective of different research fields resulted in vastly different responses. As comparing these responses could lead to a much broader answer to the research question, it was decided to include experts from a wide variety of different research fields, instead of focussing on a select research field or fields. For this reason, a specific research field was not defined as a criteria for selection. Beyond helping to refine the sampling criteria as well as the specific set up of the interviews, the trial interviews were not suitable to be included in the main sample. This is because they followed a different set up than the main sample interviews, also following a different line of questioning.

As for the actual sample selection, the researcher used several methods of finding suitable candidates. Firstly, to produce an initial list of candidates who would potentially meet the criteria and who would also be willing and available for an interview, the researcher inquired into their own extended network. This is clearly an instance of convenience sampling, however as the researcher merely used this step to form a preliminary list of potential candidates and did not select candidates on this basis only, it

does not invalidate the credibility of the sample. 17 people were contacted following this inquiry.

From these people seven eventually agreed to participate, some after a short email correspondence explaining more details about the research. These seven people made up the bulk of the main sample. Of the people who rejected the request for an interview, several made suggestions of other people to contact (snowball sampling). Based on these recommendation, another 5 people were contacted, of which, only one agreed to participate. Reviewing the eight people that were selected, it was deemed necessary to include some more people from outside the researchers extended network, who also had some more practical experience with predictive policing. To do this, a Google search was conducted, searching for people who fulfilled the two criteria described earlier. Five people were contacted based on this search, of which three were willing to participate in the research.

Of the in total eleven people who agreed to participate in the research, two were interviewed in trial interviews, as was mentioned earlier. Of these two, one was interviewed again as part of the main sample. The second expert was not available for a second interview. As such, the main sample studied in this thesis includes 10 experts. A detailed description of the sample can be viewed in appendix F. Please note, to preserve anonymity, only details relevant to this thesis were included in this sample description. Which are: a description of the experts academic background, a description of their knowledge of predictive policing (practical/academic/both), and finally how the expert was found (convenience/snowball/internet search).

As can be seen in the sample description in appendix F, this thesis has mainly studied academic experts. Some of which, gained in depth knowledge of predictive policing practice, either by doing empirical research into the police or doing practice oriented research for the police. Besides this, there are several important differences between the experts that were interviewed. Firstly, their academic background. As can be seen, most of the experts had differing academic background, which also meant different points of view from which they approached the study of predictive policing. This gave a broader basis of contextual information which allowed for a much more comprehensive view of the context of predictive policing. This also provided a comprehensive view of the multitude of different ways in which values, and their meaning in a practical context, could be approached. Among the research backgrounds, law and criminology were most represented. Secondly, the experts came from different countries in Europe. This also

gave an interesting variety to the sample which further added to the broader more comprehensive basis of contextual knowledge. Among these different countries, the Netherlands was represented the most.

3.3 Data Analysis Method

In qualitative research, “the researcher is the primary instrument for data collection and analysis (Merriam & Tidell, 2016, p.16).” This makes sense because the goal of qualitative research is understanding, and there is no tool more suitable for constructing understanding than human interpretation, as humans are able to flexibly respond and adapt their understanding to new information (Merriam & Tidell, 2016). On the flip side, it should also be emphasized that qualitative research is highly subjective and therefore subject to bias. However, instead trying to eliminate this bias completely, it is important for a qualitative researcher to make their assumptions clear and to show how these shape the collection and interpretation of data (Merriam and Tidell, 2016). This was done in the theoretical framework where two ways in which broad explainability could relate to the values identified were outlined – fostering and balancing. Furthermore, two factors of explanation, i.e. factors that help determine what the right explanation is, that this thesis will look at were also identified - (1) the intended audience and (2) the right content. To which was added (3) related values. Finally, with regards to what the potential content could be 7 components of algorithmization were outlined.

As for the actual analysis of the results, this thesis follows an inductive, rather than a deductive process. Qualitative research is often inductive as it often has to deal with a lack of theory or lacking theory to explain a phenomenon (Merriam & Tidell, 2016); as is the case in this thesis. Being inductive, means that the theoretical framework is not tested in an experiment, which would be deductive, but rather is created on the basis of what can be induced from existing literature. Which in turn, provides the context within which the researcher inductively derives findings from the data, in the form of “themes, categories, typologies, concepts, tentative hypotheses, and even theory about a particular aspect of practice (Merriam & Tidell, 2016, p.17).” In conclusion, analysis of the results from the interviews will be done based on inductive interpretation.

In order to allow the researcher to inductively interpret the interviews, all the interviews were transcribed. This was done by first using the transcription tool of Microsoft Word Web, which is a service included in the Microsoft Office 365 online version of Word. After getting a rough transcription through this method, the researcher

listened again to all the recordings and refined the transcriptions. The researcher opted for a word for word transcription of the interviews. As Patton (2015) explains, when summarizing interview responses, e.g. by removing or improving certain broken sentences/answers, one is imprinting their own bias onto the data. In order to have an unbiased basis for analysis, the researcher did a word for word transcription of the interviews, leaving intact also broken sentences. Furthermore, by doing the transcription themselves, the transcription process can actually serve as a first way for the researcher to familiarize themselves with the data before actually analysing it (Patton,2015).

To aid the researcher in analysing the data, a coding software was used which is specifically designed for qualitative and mixed research methods (MAXQDA, version 20.4.1, 2020). All the codes were created and assigned manually by the researcher. The software merely helped with the retrieval of relevant segments, based on the codes that were manually assigned, which allowed the researcher to analyse responses on a certain topic across all the interviews.

Initially, the coding followed the structure of the interview. As such, there were two main categories of codes to begin with: 'Interview Part 1' and 'Interview part 2'. Within these main categories, the first codes were created. Under Interview Part 1, unique codes were created for each of the 12 values. Then under Interview Part 2, two subcategories were created: 'Most Important Values' and 'How to Foster Them'. Within the first subcategory the codes 'Important Values' and 'Reasons for Importance' were created. In the second subcategory codes for each of the factors of explanation were created: 'Audience', 'Related Values' and, for content, 'Components of Algorithmization'.

After this initial round of coding, it was concluded that there was a noticeable difference in the way the experts approached the values derived from the discussion of the main benefits of predictive policing, and the values derived from the main concerns with predictive policing. As such, two new subcategories were introduced under the 'Interview Part 1' main category. These were: 'Values from the Benefits of Predictive Policing' and 'Values from the Concerns with Predictive Policing'. The value codes were then placed under their corresponding sub category. Additionally, it was discovered that in both parts of the interviews, all 12 values had been discussed. This resulted in a practical problem in terms of analysis, namely how the codes could be applied in a way that would allow the researcher to, (1) differentiate between the different values, and (2) differentiate between segments relevant to either part one or part two of the interviews.

This differentiation was necessary for the analysis because part one and part two of the interviews were targeted at answering distinctly different types of questions related to different parts of the theoretical framework (see interview schedule in appendix B).

Preferably, this distinction had to be made without having to create a twin code for each value, under 'Interview Part 2', and then subcodes under these for each factor of explanation discussed (Audience, Related Values, and Components of Algorithmization). As such, the problem was remedied by giving each unique value code under 'Interview Part 1', two corresponding subcodes: 'Part 1' and 'Part 2'. The benefit of this coding method, over the 'twin codes' method, is that this coding method also allowed the researcher to quickly gather an overview of all the information with regards to a certain value, from both parts one and two of the interviews, by using the value's main code, without having to select all the twin codes under the 'Interview Part 1' and 'Interview Part 2' main categories. This greatly diminished the potential for error during analysis. The complete coding scheme used can be viewed in appendix G. For the purpose of readability, the subcodes 'Part 1' and 'Part 2' under each of the value codes were collapse.

When analysing the interviews, the Retrieved Segments function of MAXQDA was used. This function allowed the researcher to retrieve segments containing one or more codes and compare these segments. For example, the researcher could compare what was said about accountability in part two of the interviews with regards to the intended audience by selecting the codes: 'Accountability', 'Part 2', and 'Audience'. Using this function the researcher was able to analyse and compare different information across all the values and different topics that were gathered in all the interviews.

Observant readers will notice that the coding scheme roughly resembles the same structure as was used in the fill in documents (appendix D) that were filled in with the experts during the interviews. The fill in documents were used during the coding and analysis as a reference to the basic structure of the experts responses. However, they were not considered as main sources and were merely considered as complementary to the transcriptions. As such, they were also not coded themselves. One exception on this rule has to be mentioned however. During one of the interviews, the recording equipment failed and for that instance, the researcher relied on the information in the fill in document to supplement the lost information – in this instance extra notes were added at the end of the interview when the failure was discovered. This is in line with what Patton (2015) recommends with regards to notes taken during interviews.

4 Results

4.1 General Comments on Explainability

Before going into the specific results for each of the values, it is important to give an overview of the general comments that the experts made which are relevant for our understanding of technology-centred explainability and broad explainability. Firstly, several comments were made that are relevant for the feasibility of technology-centred explainability. Expert 4 and 2 commented on the difficulties in realizing explainability due to the complexity and opaqueness of the technology. Expert 4 explained that, in the case of the Dutch police that she researched, intelligence officers were placed between the predictive algorithm and the police teams acting on the predictions. It was the job of these intelligence officers to judge the predictions and provide context to them so that they would be comprehensible for the police teams. However, often the logic behind these predictions is too complex for the intelligence officers to understand, which meant that these intelligence officers would substitute this information with their own suggestions. Which, according to her, is problematic because you ask the police teams to act on a new insights but you cannot substantiate how these insights were formed based on the system.

Expert 4 also explained that this opaqueness of the system is necessary, according to the developers. According to this group, the opaqueness is the strength of the system, because if we understood how it generated new insights then these would not really be new insights. Expert 4 thus concluded that there is a difference between these two groups in how explainability is valued – *“the developers say: you need to have this lack of explainability to get these new insights. While the translators and the users say, yeah, but I need explainability to be able to use it.”*

Expert 8 also made some comments from the perspective of development. According to her, explainability really depends on the design of the system. If a system uses a very broad range of data, then explainability could actually be helpful for citizens and police officers because it can give more insight into the context of a prediction which could lead to insights into criminal behaviour. Furthermore, she argued that if a system could provide counterfactual explanations than this would even be better, because then the police would be able to understand why it is better to take one action over another. This is all of course dependent on whether the system is designed in a way that these types of explanations can be given.

Secondly, there were also comments made relevant to our understanding of the feasibility of broad explainability. Expert 10 was most adamant on this point, making several comments with regards to this. To start, he addressed the fact that many people involved in the predictive policing process have no idea about how it works. As an example, he clarified that many people in the judiciary are completely unaware of new technologies and how they work, and that they would need a lot of basic explanation on computers and statistics for them to even understand explanations with regards to the practice. Then they also stressed that it would be hard to estimate which types of information are relevant for different audiences, and that this is even more true for the general public. Basically, it is very hard to determine how much a certain audience already knows about a certain topic and, with that, whether an explanation will be comprehensible for that particular audience.

A second point that they made is that police departments might be reluctant to share information with regards to their policing practice and as such, it will be very hard to practice broad explainability. As Expert 10 clarified, *“some of the ways the system works are implicit and cannot be made explicit for a variety of reasons. Because some people are going to lose face to face if it's explicit or because some of the practices are illegal or we're not totally sure if it's legal, so it's better not to mention them.”* To this they also later added: *“some of the ways that the system works are probably contradictory with some other ways that the system works.”* And eventually they concluded: *“You're going to run into some of these contradictions, or some of these implicit rules that cannot be made explicit, and as such, if you try to reach broad explainability, you're going to face resistance from within the organization.”*

Police departments might not even use predictive policing with the goal to reduce crime rates, they might have acquired it for different reasons. To argue this point, Expert 10 explained that police departments throughout Europe are sometimes heavily underfunded and that, by claiming to use predictive policing, they can get budget from the government to build, for example, a new data management system. In which case, the goal of having the predictive policing tool is not to reduce crime but actually to get budget for other related things.

Finally, Expert 10 also made several points with regards to how and when the explanation is actually given, which are relevant to whether they could actually foster any kind of value. First, they argued that explanations should be given honestly. With this they meant that if an explanation is given it must reflect the actual truth of the practice

and it should not be portrayed better than it actually is. For example, an organization should be honest about the level of expertise that the people working with the system have. They doubted however, the extent to which organizations would be willing to be completely honest, for the aforementioned reasons. Secondly, they argued that if you want organizations to give honest explanations, then they must have a system that functions properly, otherwise they might be inclined to hide certain aspects. This is especially important for values such as trust and fairness. Although honest explanations about systems that do not function well could also foster trust. Then finally, they also argued that people within the organization must have a mission to accomplish certain objectives, e.g. making the police more efficient or effective. To summarize, across the various interviews, general comments were made on the feasibility of (broad) explainability in the context of predictive policing. These are relevant to be discussed at a later point as these are indicative of certain conditions that have to be met before values could potentially be fostered with (broad) explainability.

4.2 Results for each Individual Value

Here, the specific results for each of the twelve values that were identified as being at stake in the context of predictive policing will be outlined, based on the results gathered in the interviews. Each section will start with the definition of the value which was determined in this thesis. These definitions were also shared with the experts before the interviews (see appendix C) and shown to the experts during the interviews (see appendix D). The discussion of each value will follow a common structure: (1) any comments relevant to the definition of the value will be discussed if necessary; (2) the relationship with broad explainability as a practice will be discussed; and (3) the factors of explanation – audience, related values and components of algorithmization – will be discussed. Each subsection shall end with a table displaying the main takeaways per value, with regards to each of the four discussion points identified in the research background. Please note that when discussing the related values, not only the comments made in the second part of the interviews will be considered, relevant comments made by any of the experts from the first part of the interviews will also be included.

4.2.1 Values Resulting from the Benefits of Predictive Policing

4.2.1.1 Efficiency

Efficiency was defined in this thesis as: *the extent to which police resources are optimally allocated in spatial and temporal sense*. When it comes to the definition of

efficiency, there were no significant objections from the different experts. However, some experts had questions with regards to the inherent difference between efficiency and effectiveness. Expert 6, for example, explained that, in Norwegian, these two concepts are not separated but are actually one and the same word. Expert 2 agreed that the relationship between these two concepts resembles the paradox of the chicken and the egg – which one precedes the other? Furthermore, Experts 2 and 3 also mentioned efficacy in tandem with efficiency and effectiveness. According to Expert 2, efficacy pertains to the ability to produce desired results.

Most experts agreed that efficiency, often in tandem with effectiveness, were very important in the context of predictive policing. Experts 4 and 5 pointed out that efficiency is often the reason why the police adopts predictive policing tools into their practice. Expert 4 said that, in the case of the Dutch police, predictive policing was seen as a method of making better use of the scarce resources that the police has. Expert 9 even made the claim that, when it comes to the legitimacy of predictive policing from the standpoint of the general public, efficiency and effectiveness are important metrics for the police to communicate to the public when explaining why they are using it.

In terms of how efficiency relates to broad explainability, most experts agreed that giving explanations about the entire practice of predictive policing could potentially increase the efficiency of policing practice. Experts 1, 2, 3 and 5 argued that explaining how the system works as well as how the practices around the system work, could potentially allow the police to make better use of the system. To illustrate this they pointed to the fact that many police officers do not understand how the system works and are therefore reluctant to use it. Expert 9 followed a similar line of reasoning, adding that explanations about the predictive policing practice could foster acceptance among police officers, which could make it more likely that they would follow the predictions of the system. Which could then potentially lead to a more effective use of the predictions in policing practice.

Experts 4 and 8 made a different argument but in the same direction, namely that, if one can explain to police officers why a certain prediction was made, it can help them understand why they are being sent somewhere which can make them better utilise the information from the predictions; which could make them more efficient. This is important, according to them, because the goal of predictive policing is crime prevention, and in that regard, police officers are not sent somewhere to catch a criminal in the traditional sense. More often, officers are sent somewhere to deter criminals from

committing crime. However, this means that police officers have to change their approach to policing and for that they need to understand the context of the prediction so that they can better understand how to realise this deterrence. Expert 6 took again a different approach and argued that if one can give an explanation that can show that the police works efficiently because of the system, then that increases the perception of efficiency.

The fostering relationship between broad explainability and efficiency was however not taken for granted. As Expert 7 addressed, some people see the relationship between explanations and efficiency as a zero sum game. He however doubted this assumption and argued that giving explanations does not necessarily have to interfere with efficient crime prevention. Expert 10 addressed this point by stating that efficiency could be fostered on the condition that we accept the opportunity costs of explaining things. Expert 3 and 8 addressed the fact that explanations could be given to the general public as well, in which case it could be that criminals – who are inherently a part of the general public – would adapt their behaviour in an attempt to cheat the system. In which case, broad explainability would have to be balanced against efficiency. Expert 3 added however, that she did not think that this would necessarily be the case and furthermore, Expert 8 added that giving police officers explanations that would help them understand the context of the predictions and why they were given could help them anticipate changes in criminal behaviour.

This forms a natural bridge to part two of the interviews. Three experts (2, 3 and 10) discussed the value of efficiency in this part of the interview. In terms of audience, besides the police themselves, the experts named the general public and decision makers as important audiences for explanations fostering efficiency. As Expert 2 argued, explaining about the predictive policing practice might create some societal pressure for the monitoring of efficiency. In other words, it might incentivise the police to pursue efficiency because they know society is watching. Experts 3 and 10 stressed the need for decision makers and auditing bodies to receive explanations as well. This could lead to more efficiency as it is these parties that have to monitor the police and who have the power as well as the incentive to improve their efficiency.

In terms of related values, of course effectiveness was often named as most of the expert assumed an inherent relation between these two values. Then also accuracy was named by Expert 7 as something that could improve the efficiency. Furthermore, comprehensibility was named as a related value because if predictive policing was made comprehensible it could enable the police to work with the system more effectively.

Finally, accountability was mentioned for the reasons described previously, with regards to the idea that evaluation would put pressure on the police to work more efficiently.

The components of algorithmization that were mentioned as being important in explanations fostering efficiency were: monitoring and evaluation, socio-technical relations and expertise. However, Experts 10 and 3 actually named all the components as important when explanations are given for the purpose of scrutiny. Expert 10 named expertise as being important, however not as something that needed to be explained but as something that needed to be improved within the police organization in order to foster more efficiency.

Relationship to Broad Explainability	Audience	Related Values	Components of Algorithmization
- Predominantly fostering relationship. if explanations are given to the police themselves. - Could be balancing if opportunity costs of explanations are too high.	- Main audience is police themselves. - Could also be general public and decision makers.	- Closely related to effectiveness. - Accuracy and Accountability could be helpful. - Comprehensibility is necessary.	- Monitoring and Evaluation, Socio-Technical Relations and Expertise. - All of them could be important.

4.2.1.2 Effectiveness

Effectiveness was defined as: *the extent to which police operations are successfully reducing crime rates*. As was discussed, the definition of effectiveness was questioned with regards to its relationship with efficiency. As such, most experts said that the relationship between effectiveness and broad explainability could be fostering, based on the same arguments that were made in terms of efficiency. As with efficiency however, the experts did not take the fostering relationship between broad explainability and effectiveness for granted. Expert 1, looking from the perspective of criminal law, made the argument that combining explainability with effectiveness seemed to be illogical, especially when considering whether explainability could foster effectiveness. This is because, in criminal law, the effectiveness of a system is already established and is part of what the judiciary scrutinizes during criminal proceedings. As such its illogical to think about explanations possibly fostering effectiveness.

Furthermore, some comments were also made with regards to actually being able to measure the effectiveness of predictive policing. As Expert 4 explained, predictive policing is concerned with preventing crime, in which case it means that crime does not take place. This automatically leads to the problem of how one could possible outline all the crimes that were not committed due to the police's use of predictive policing. This is

inherently impossible as one cannot make any definitive claims with regards to things that never happened.

Finally, Experts 4 and 7 explained that predictive policing in Europe mostly consists of place based predictive policing with the aim of preventing crimes such as burglaries and car thefts. Increasing policing efforts focussing on those crimes might cause criminals to choose different avenues of crime such as cybercrime. Also, as Expert 7, stressed that the dangers of criminals cheating the system becomes more problematic when talking about effectiveness, as criminals could adjust their behaviour to seem less professional. Which, theoretically, would diminish the effectiveness of person based predictive policing which focusses on identifying repeating offenders. Expert 8 drew the fostering relationship between broad explainability as a practice and effectiveness even further into question, arguing that having to allocate precious police resources to explaining policing practice could interfere with police actually being able to effectively reduce crime.

Three experts (7, 9 and 10) discussed effectiveness as an important value to foster in the context of predictive policing. In terms of audience, they argued that explanations fostering effectiveness would have to be targeted internally at, for example, police officers. They argued that the public would not have an impact on how effective the police is. In terms of the related values, of course efficiency was mentioned very often, and accuracy and security were also mentioned. Efficiency and accuracy as helpful for fostering effectiveness. With regards to security, Expert 9 argued that the goal of predictive policing is to prevent crime in order to keep people safe, and that having a more effective police would in turn imply an increase in security. Besides these values, there were no other values named as being important to the process of fostering effectiveness, however, effectiveness itself was mentioned as being important for fostering other values such as trust, responsibility, fairness and accountability but in these instances as something that would needs to be showcased or proven in order to foster these values.

Finally, the components of algorithmization that the experts believed are needed to foster effectiveness are: technology, organisational structure, socio-technical relations and expertise. These were mentioned as the experts believed that, to foster effectiveness by targeting explanations at people within the police organization, it is important to give them an account of both how the system works and how they are supposed to work with it. Expertise in this sense would be necessary to be explained in order to foster trust with

the police officers. As Expert 9 explained, police officers on the street rely on information officers to translate the predictions into actions. Explaining the police officers on the street about the level of expertise of the information officers, could result in them trusting the predictions more, as they would know that qualified people are working with the system, which, in turn, could cause them to follow the predictions of the system.

Relationship to Broad Explainability	Audience	Related Values	Components of Algorithmization
- Fostering for the same reasons as efficiency. - Effectiveness of predictive policing is hard to prove, so the impact of explanations is hard to determine.	- Main audience is police themselves.	- Closely related to efficiency. - Accuracy and efficiency could be helpful - Could benefit Security. - Important for Trust, Responsibility, Fairness and Accountability	- Technology, Organisational Structure, Socio-Technical Relations and Expertise.

4.2.1.3 Accuracy

Accuracy was defined as: *the extent to which a predictive algorithm makes correct predictions*. None of the experts had any specific comments on how accuracy was defined. Many of the experts thought that broad explainability as a practice could have a fostering relationship with accuracy and they gave a few different arguments as to why this relationship could be fostering. Experts 2, 3, 4, and 8 all made similar arguments, claiming that broad explainability as a practice would mean that more is known about the exact variables that lead to a prediction, which would help the police identify those variables that are actually relevant to predicting a certain type of crime and which are not. This in turn, could help improve the accuracy of the predictive policing algorithm. Expert 4 gave a good example, she explained that at the Dutch police, their algorithm predicted a higher chance of car thefts in a park which was inaccessible for cars. This was obviously an inaccurate prediction and by practicing explainability, people could understand how an algorithm could make such an inaccurate prediction and remedy this. As Expert 3 said: *“you know what it is supposed to do, and then you can stop when it doesn’t work as it is supposed to.”*

Note however, that these experts were only considering the technology, which is understandable when considering the way accuracy was defined in this thesis. With regards to broad explainability, Expert 8 made an interesting argument. She said that by having broad explainability, people would also be able to better understand what kind of police action was effective and how criminal behaviour changed based on police actions

taken. In turn this would allow developers to feed this information into the algorithm, hence making the system more accurate. Put differently, having broad explainability could help the police understand better the crimes they are trying to predict, the actions that they can take based on these predictions and how effective those actions were in the past, which they can then use to improve the accuracy algorithm itself.

As opposed to the experts discussed up until now, Experts 7 and 9 did not see a relationship between giving explanations and fostering accuracy. Expert 7 made two arguments regarding this. Firstly, he argued that the extent to which predictive policing practice is accurate is dependent on how predictions are put into practice, which tends to run into resistance among street level police officers. As such, he argued that, to the higher ups in the police department, the accuracy of predictions does not seem to be as important as actually getting street level police officers to use the predictions. Secondly, he also pointed out that accuracy is very hard to prove. As he explained, this is for the same reason as with regards to effectiveness. Namely that, because predictive policing is aimed at preventing crime, you never really catch any criminals, and therefore there is no way of proving that a predictive algorithm actually correctly predicted a crime.

Only Expert 1 argued that accuracy is important to foster in the context of predictive policing. In terms of audience, he named the judiciary, the police themselves and the general public. However, he did not provide any arguments on how providing them with explanations could foster accuracy. Rather, he argued that it is important to give them explanations about accuracy because of the other values that can be fostered by doing so, which was mainly trust according to him.

With regards to the related values, of course effectiveness and efficiency were seen as related values that could benefit from increased accuracy. Expert 1 argued for trust stating that if one can show that a system is truly accurate, then people would automatically trust it more. Expert 4 also linked accuracy to trust, arguing that accuracy is a value that needs to be both realised and explained when fostering trust. Finally, Experts 3, 6, and 7 also linked accuracy to fairness. Expert 3 explained that many of the fairness related problems come from the fact that an algorithm is not accurate for everybody. Expert 6 argued that when it comes to fostering fairness it's important to know more about the accuracy. Why is explained by Expert 8 who argued that accuracy is necessary to determine whether a practice is fair. In this same line of argumentation, Expert 5 grouped accuracy, effectiveness and efficiency together as being values that exemplify better or improved policing practices. Expert 1 only named the technology as

component of algorithmization that needs to be included in an explanation fostering accuracy.

Relationship to Broad Explainability	Audience	Related Values	Components of Algorithmization
<ul style="list-style-type: none"> - Fostering as it would allow the police to identify inaccuracies in their predictive algorithm, as well as practices that were successful in the pas. - Accuracy is hard to prove, similar to Effectiveness. 	<ul style="list-style-type: none"> - Main audience is police themselves. - Could also be general public and judiciary. 	<ul style="list-style-type: none"> - Effectiveness, Efficiency, Trust and Fairness could be benefitted. 	<ul style="list-style-type: none"> - Technology.

4.2.1.4 Security

Security was defined as: *the extent to which predictive policing practice keeps people safe from risks arising from potential crime*. On the definition of security, Expert 1 made the argument that security should also include the security of the system itself, and in that sense, he argued that giving explanations about how the system actually functions could enable people to hack the system, making the predictive policing practice less secure in that sense. Expert 7 also made a comment on the definition of security, stating that whether security can be fostered through broad explainability as a practice is dependent on how security is defined and that security was defined rather narrowly here. However, he did think that security could be fostered through broad explainability as it could foster a trust relationship between the police and the public which could make them feel more secure. He also argued that explanations could help police officers follow predictions with greater confidence as they understood better how predictions were made, which could lead to an increase in security as well. Expert 9 argued that security is the ultimate goal of predictive policing.

Expert 2 and 9 made arguments with regards to the idea that security is not an absolute thing but in many cases comes down to a feeling of being secure. But where Expert 2 saw this as a cause for questioning whether there was a relationship between security and broad explainability, Expert 9 actually argued that broad explainability could only have an impact on security as a feeling and not on security in absolute terms. He argued that knowing about all the ways in which predictive policing practice is attempting to make the world safer could enhance people's sense of security.

In terms of how broad explainability relates to security, some of the experts claimed that this could be fostering. Expert 3, claimed that security could be fostered through broad explainability, although she recognised that sometimes, information should be withheld from the public. She counterargued this by stating that in practice, withholding information should be an exception that has to be properly justified rather than a rule. Expert 4 argued that explaining about predictive policing practice could help the police understand the patterns in crime better and as such make people more safe from potential crime. Experts 5 and 6 go a step further and they argued that by giving citizens explanations, they could be made an active part in the crime prevention efforts of the police because knowing about the risks of potential crime, how these predictions are created and what is done with them will also allow citizens to take appropriate action to protect themselves. Expert 10 however, provided a counterargument to the idea that security could be fostered through explanations and this was that this is wholly dependent on whether predictive policing in and of itself enhances security. For this he argued, there is no definitive proof yet.

Only Expert 5 chose security as an important value to be fostered in the context of predictive policing. In terms of audience, she argued that explanations targeted at both the police officers as well as the citizens could potentially foster security. Concerning related values, Expert 5 argued that privacy is related to security because, within the context of predictive policing, the pursuit of security always has to be balanced against the value of privacy. She also made a link to trust arguing that if predictive policing increases security, this would also foster more trust. Which is in line with what Expert 8 argued as well. Lastly, she also argued for a link with explainability because this could allow citizens and the police to better use predictive policing to improve security. Security was also linked to efficiency, effectiveness and accuracy. For example by Expert 9 who argued that a more effective police would mean more security.

Finally, Expert 5 argued that all the components of algorithmization are important for fostering security through explanations. The reason for this is that, all of them need to be explained to citizens and police officers so they can understand how their actions in relation to the predictive policing tool could improve policing practice and prevent crime, or in that sense increase security.

Relationship to Broad Explainability	Audience	Related Values	Components of Algorithmization
<ul style="list-style-type: none"> - Could be fostered if explanations allow police to perform better. Or if explanations enable citizens to protect themselves. - Could only foster sense of security with citizens. 	<ul style="list-style-type: none"> - Both police officers and citizens. 	<ul style="list-style-type: none"> - Efficiency, Effectiveness and Accuracy could be helpful. - Trust could be benefitted. - Privacy could be endangered. 	<ul style="list-style-type: none"> - All of them could be important.

4.2.2 Values Resulting from Concerns with Predictive Policing

4.2.2.1 Explainability

Explainability as a value was defined as: *the extent to which information can be communicated in a way that can be understood by recipients*. On the definition of explainability, each of the experts needed some additional explanation and information about the difference between explainability as a value and broad explainability as a practice. Which also meant that the question of whether explainability as a value could be fostered through broad explainability as a practice was especially hard to discuss during the interviews. Many of the comments made by the experts were actually on whether broad explainability and explainability as a value could actually be achieved. Which were discussed at the beginning of this chapter on results. Fortunately, some of the experts were able to make some arguments on the topic.

Expert 3 argued that if one is forced to give explanations, one is also motivated to improve the explainability of what they are trying to explain. Expert 4 gave an elaborate explanation of where the need for explanation within the police lies, in the end arguing that, by providing more context through broad explainability, police employees would be better able to understand policing practice and their role therein. Which, in other words, means that predictive policing practice becomes more explainable to them. Expert 6 made a similar argument, claiming that broad explainability would improve people's understanding of the context of predictive policing practice which would also enable them to understand explanations about this practice better in the future. Expert 7, 8 and 9 also made similar comments about how more knowledge about context would help people understand future explanations better as well as enable people to provide better explanations in the future. One important caveat to this argument however, was mentioned by expert 10, which is that people could be overwhelmed with explanations of things that are not relevant to them.

Half of the experts (1, 3, 4, 5, 6) chose explainability as an important value to foster in the context of predictive policing. With regards to the best audience to target in order to foster explainability, all five experts recognised that everybody in the process of predictive policing would be able to understand future explanations better by receiving broad explanations. Expert 3 made an additional point on audience that is interesting to mention, she argued that for all values, there are several points at which explanations should be given. First, the developers need to give explanations to the users/the police so that they know how to use the technology. After which, the users/the police have to give explanations to authorities who have to scrutinize their use of the technology.

In terms of related values, explainability was often linked to transparency and comprehensibility. However, the experts provided many different explanations on how these concepts exactly relate to each other. Expert 1 argued that you need transparency and comprehensibility in order to ensure explainability. Expert 4 also named comprehensibility as important for explainability. Expert 2 on the other hand argued that if you have transparency without comprehensibility and explainability then it would be fake transparency because no one would be able to understand it. Besides transparency and comprehensibility, explainability was mentioned as being important for trust and responsibility. Expert 1, 5 and 6 argued that in order to have trust a process should be transparent and explainable. With regards to responsibility, expert 7 made the most elaborate argument, arguing that broad explainability can help people understand that using a predictive policing system comes with a certain responsibility. In other words, it can heighten peoples sense of responsibility as well as allow them to recognize when the system is being used in an irresponsible way. For example, when the system is discriminating against certain neighbourhoods. Finally, explainability was argued to be fundamental for accountability, which will be explained in the next section on accountability.

It should not come as a surprise that those experts who discussed explainability in part two of the interviews argued that all of the components of algorithmization should be included in an explanation. Expert 4 actually did point to some components as being more important than others, which were: technology, information relations, and socio-technical relations. She explained that the technology and information relations are important because a basic understanding of the system and the data that is being put in is necessary, and socio-technical relations are especially important because this addresses how the system is and can be used in the broader socio-technical context.

Relationship to Broad Explainability	Audience	Related Values	Components of Algorithmization
<ul style="list-style-type: none"> - Could be fostered as more understanding of the specific context would allow for a better understanding of future explanations. - Might need to be balanced if people are overwhelmed with information 	<ul style="list-style-type: none"> - Everyone inside and outside the police organization. 	<ul style="list-style-type: none"> - Transparency could be preconditional or conditional. - Trust, Responsibility and Accountability might be benefitted - In the end all values could potentially benefit from Explainability 	<ul style="list-style-type: none"> - Technology, Socio-Technical Relations, Information Relations. - All of them could be important.

4.2.2.2 Accountability

Accountability was defined as: *the extent to which the use of predictive policing algorithms can be assessed and consequences imposed based on those assessments*. All the experts agreed that broad explainability could definitely improve the accountability of predictive policing practice. Taking the perspective from inside the police organization, Expert 6 argued that explaining people the kind of data that is put into the system and how the algorithm is trained could foster accountability. Expert 3 argued something similar, explaining that people need to understand how the system works in order to be accountable. However, some of the other experts stressed the point that explaining the technology alone would not be sufficient to foster accountability. Expert 1 explained that if we only provide an explanation consisting of mathematical equations then this would obviously not foster accountability. In his words, from the criminal law perspective, *“If I step in the role of criminal defence lawyer, for example, if I give him a complete explanation with mathematical formula and, with all the math or historical data and all the explanations on how the machine learning model was built, how it was trained and everything else. Well he would just look at me, surprised.”* Expert 9 argued that the police is still accountable in the end, and stressing the need to look at the whole predictive policing practice when holding the police accountable.

It should be clear that these comments relate to technology-centred explainability. The experts were much more positive about broad explainability Expert 5 argued there is a need for broad explainability because accountability issues could rise in all stages of the process, and not only with regards to the technology. Expert 4 argued that if you are able to explain the whole predictive policing practice, then you *“Get away from this idea: there's either the AI predictions or the human decision's, but it's more like who is involved in which part of the process of using AI.”* Expert 8 also stressed the need to explain the whole practice in order to be accountable *“I think you really need explanations to even be*

accountable, because if you cannot explain what you've been doing or why, then it's not accountable."

Seven of the ten experts that were interviewed (1, 2, 3, 6, 7, 9, 10) chose accountability as one of the most important values to foster in the context of predictive policing. Of these seven, Experts 1, 3, 6, and 7 said that people within the police organization would be a target audience for explanations fostering accountability. Expert 1 argued that more detailed explanations would have to be given to the top of the policing hierarchy as they have the power to hold people accountable. Almost all the experts also concluded that explanations could be given to the public, Expert 3, for example, argued that society could hold the police accountable. In the words of Expert 8, *"In order to create accountability. There needs to be someone claiming the account and someone giving the account right and so in order to claim an account you need to have at least a baseline of information about what's going on."* Expert 10 stated that we normally look at the judiciary when it comes to holding people accountable but that in the case of predictive policing the judiciary might be too busy establishing basic facts to be able to really delve into the particulars of predictive policing and therefore another appropriate audience could be journalists.

Many of the other values were named as being related to accountability. Experts 7 and 8 claimed that explainability is fundamental to accountability. As Expert 7 stated: *"To give explanations is the basic principle that underpins the idea of accountability."* Similar arguments were made for responsibility, and transparency, where the experts argued that these are needed in order to have accountability. Expert 3 for example, argued that to be accountable you first need to be responsible, and Expert 2 said that responsibility is a precondition for accountability because *"If you have a clear division of responsibilities and then you know who to call to be accountable."* Expert 9 explained that you need transparency to be able to assess something. Expert 2 also said that transparency is important for accountability, but argued the other way around. She said that *"If you're accountable then, transparency follows."* Expert 6 argued this way for both responsibility and accountability – saying that that if you have accountability, then this would also foster responsibility and transparency.

Some experts also argued that accountability is related to trust and fairness. Expert 2 argued that if you want to be fair, then you need to be accountable for the underlying actions as that would make you ensure that the processes are fair. Later however, she argued that fairness comes before accountability as a precondition. Expert 9 also named

fairness as related to accountability. In terms of trust, Expert 10 argued that by holding people accountable, you create trust in the institution, and Expert 6 also argued that accountability would foster trust. Expert 7 was even stronger in his argument and argued that *“Accountability is a building block for trust.”*

Expert 4 argued that explainability, accountability, responsibility, transparency and trust are all interrelated. She said that *“I think that you need this explainability to foster the responsibility because you need to understand how the decisions are made for example to know what you are responsible for.”* To which she later added *“ I think responsibility and accountability are kind of linked, so if you know where you're responsible, you also know where you're accountable.”* But on this point she did note that she did not know which one came first. On transparency expert 4 explained that *“the more transparent the rules, for example, or the decisions that are made, the better you understand your responsibility in the process.”* Finally, according to expert 4, there is also a link to trust because *“if it's explained to you how you are accountable in the use of these systems or how you are responsible and use these systems. That influences how you trust its outputs. Or how you trust that it will be okay if you use the outputs.”*

Finally, Expert 9 also made a case for privacy as a related value, he explained that accountability, together with responsibility and transparency, are important for privacy because these need to be in place on order to be able to assess the privacy. And expert 1 argued for accuracy as a related value but more in the sense that accuracy is something that needs to be explained in order to be accountable.

With regards to the components of algorithmization that would need to be explained in order to foster accountability, each expert named a number of them, but overall, each component of algorithmization was named as important to explain. Expert 3 and 6 even explicitly said that all of them need to be explained in order to foster accountability. Expert 10 provided a reason why all of them are important, explaining that *“you need to know whether or not the technology works, and then you need to know how the decision is taken based on this technology.”*

Table 4.6: Main Takeaways – Accountability

Relationship to Broad Explainability	Audience	Related Values	Components of Algorithmization
- Could be fostered either because those holding the police accountable will receive better explanations.	- Police - General public - Journalists	- Explainability, Comprehensibility, Transparency, Responsibility, Fairness, and Accuracy, could be necessary.	- All of them could be important.

		- Transparency, Responsibility, Fairness, and Trust, could be benefitted - Effectiveness, could be endangered.	
--	--	---	--

4.2.2.3 Responsibility

Responsibility was defined as: *the extent to which the duty of care for the proper use of the predictive policing algorithm has been clearly allocated*. When it came to how responsibility was defined, Experts 2 and 3 both argued that the definition should be broader because responsibility is not only about having the duty of care allocated but also about acting in a responsible way. Besides these points, several experts explained the problems with responsibility in the context of predictive policing. Experts 1, 2, and 3 all explained that when it comes to the police's use of algorithms, it's hard to pinpoint where responsibility exactly lies which sometimes means that responsibility is placed on the AI system itself. Expert 4 argued that responsibilities should be clear before the predictive policing system is actually used. But that responsibilities are often described in very abstract ways, e.g. the developer develops and the translator translates, and that these responsibilities become blurred once the predictive policing system is actually used.

Many of the experts did seem to think that broad explainability could foster responsibility. Experts 4 and 5, argued that broad explainability could help people understand where their responsibility lies and what they are actually responsible for when the predictive policing system is used. In Expert 4's own words: *"I think the responsibilities already should be clear beforehand, but should be kind of tracked along the way in the implementation and use of AI. To kind of also uncovered the unexpected things that happen in the whole practice."* Expert 5 also stated that explainability can help the police understand how to use the predictive policing system responsibly, and Experts 7, 8, and 9 also made arguments in this direction. Expert 8 agreed that explanations are needed to allocate the duty of care and Expert 9 summarized it as: *"If you explain how your predictive policing practice works, then it also becomes evident who is responsible when for what."* Expert 7 gave a more elaborate argument and explained that broad explanations could help make people aware that a predictive policing tool does not relieve them of their responsibilities but that using a predictive policing tool comes with responsibilities of its own. For example, making sure that there is no discrimination or that people do not feel threatened by the police's use of predictive policing tools. Expert

6 also added an interesting point, namely that broad explainability could also help people identify a breach of responsibility.

Only two experts (4 and 7) picked responsibility as an important value to foster in the context of predictive policing. In terms of audience, Expert 4 argued that everybody involved in the predictive policing process should be given an explanation. In terms of related values, expert 7 argued that explainability underpins the value of responsibility. Besides this, accountability, transparency and trust were named as being related to responsibility. How the experts argued that these values are related to responsibility has already been addressed in the section on accountability. However, one additional comment was made by Expert 7 on the topic of how responsibility relates to trust, and also fairness. He stated that: *“I think in general they would be rather enabled by responsibility. Right, so enhanced responsibility would then also mean that there’s potentially enhanced trust and enhanced fairness.”*

As for the components of algorithmization, Expert 4 named technology, expertise, organisational structure, organisational policy and socio-technical relations. Expert 7 also named information relations, which means that, in the end, all components of algorithmization were argued for as important. Expert 7 also explained that these components are important because, having to explain all of these things to the public also forces the police to internally take responsibility.

Table 4.7: Main Takeaways – Responsibility

Relationship to Broad Explainability	Audience	Related Values	Components of Algorithmization
<ul style="list-style-type: none"> - Could be fostered either because the police would better understand how they are responsible, and how to take responsibility/act responsible. - Could also help the police identify a breach of responsibility. 	<ul style="list-style-type: none"> - Everybody in the predictive policing process. 	<ul style="list-style-type: none"> - Explainability, Comprehensibility, Transparency, and Accountability, could be necessary. - Accountability Trust, and Fairness, could be benefitted - Effectiveness, could be endangered. 	<ul style="list-style-type: none"> - All of them could be important.

4.2.2.4 Transparency

Transparency was defined as: *the extent to which information with regards to the whole predictive policing practice is made available*. All experts agreed on the point that broad explainability could foster transparency, and most of them gave arguments pointing to the idea that transparency and explainability as a value are very closely related. Indeed, when discussing transparency, it was revealed just how much they actually relate, both as

a practice and a value, and how, in many ways, they are thought of as being inherently linked. Expert 3 for example, argued that explanations are a part of being transparent. She explained that the police has to give explanations and that this is an important element of transparency from a data protection perspective. From this perspective, people have the right to be told that their data is being used and they have the right to be explained what is being done with their data; which is all a part of transparency. Expert 6 argued that broad explainability will definitely foster transparency “*Because it will be more visible and more available the different aspects of the police model.*”

Expert 9 argued that broad explanations inevitably improve the transparency of the predictive policing process. When asked if explainability automatically implies transparency, he argued that just making data available does not automatically increase transparency, and that, what is made transparent should actually be adapted to the specific audience, so that it can be made sure that they can understand it and that they can actually use the information. Expert 5 summarized it by saying “*Of course, explainability is fostering, because if it's unexplainable it's not transparent.*” She also went on to explain that, a case could be made for transparency as being a fundamental value to all the other values because transparency is needed in order to prove when other values are being violated.

Expert 8 explained that explainability is sometimes used as a method of hiding certain things that the police does not want the public to find out. As an example, she explained that if the data that is being fed into the system is based on observations from biased police officers then that would mean the system would be biased. But by focussing on explaining how the system works, the police can actually hide this fact from the public. This of course only applies to technology-centred explainability and when asked if it would be different in the case of broad explainability, she agreed that this would of course foster transparency. Finally, Expert 4 explained that predictive policing is not only the output but also the input and that there is a lot going on before action is actually being taken based on the predictions made by the algorithm. As such, she concluded that practicing broad explainability, and making the decisions and processes surrounding the predictions explicit, would certainly foster the transparency of predictive policing.

Five experts argued for transparency (2, 4, 5, 6 and 9) as one of the most important values to foster in the context of predictive policing. The main target groups for explanations that were named, were the people in the predictive policing process and the public. However, when explaining why these audiences should be targeted, the expert

mostly discussed why transparency was important for these audiences and not how explanations towards these audiences could foster more transparency. In terms of related values, obviously explainability was mentioned along with responsibility, accountability, privacy and trust, which were explained before. Besides these, also comprehensibility and fairness were named. Comprehensibility was named for example, by Expert 1. However, they did not specify how comprehensibility relates to transparency. Fairness was named by Expert 4, who argued that the process need to be transparent in order to determine if it was fair.

In the context of discussing transparency, expert 7, made an interesting general comment on how values such as trust, accountability, responsibility and transparency relate to values such as effectiveness. He explained that: *“I think when we speak about transparency. It's kind of between the idea of fostering trust and accountability and responsibility. But then, on the other hand. You know we spoke about the possible dangers that you know, explaining things a bit too much or too openly could undercut effectiveness. So I think it kind of sits in that gulf, between on the one hand concerns about effectiveness and on the other hand on the benefits for transparency, for responsibility and accountability. So I would say it's probably a balancing thing.”* Which he confirmed to mean – giving explanations is already included in transparency to some extent, but if you are too transparent then that might go at the cost of effectiveness.

With regards to the components of algorithmization, the experts argued that all of these are important to explain. In terms of explaining them internally, a complete explanation could actually help foster an organisational culture in which transparency would be interwoven, according to Expert 2. Expert 6 made a similar argument, namely that explanations towards managers and decision makers would allow them to understand better how to make processes more transparent.

Relationship to Broad Explainability	Audience	Related Values	Components of Algorithmization
- Directly fostering relationship. - Explainability could be considered a part of Transparency.	- Both people within the predictive policing process and the general public.	- Fairness, Trust, Privacy, Accountability, and Explainability, could be benefitted. - Effectiveness, could be endangered.	- All of them could be important.

4.2.2.5 Comprehensibility

Comprehensibility was defined as: *the extent to which information about predictive policing practice can be understood by the recipient.* Comprehensibility was another value that some of the experts regarded as inherently linked to explainability. Expert 5 argued that *“If it’s not possible to explain it, you cannot understand it.”* Furthermore, when asked if broad explainability could foster comprehensibility, most experts responded affirmatively. For example, Expert 3 said: *“I suppose in practice we could say that they are very much related. If you give an explanation that is not comprehensible, is that even an explanation?”* And Expert 7: *“If you explain things then obviously this positively impacts the ability to comprehend them.”* Expert 6 argued that giving broad explanations fosters comprehensibility because *“Then they will also have to explain more about the context.”* Which was also Expert 4’s argument, to which she added: *“The more you explain the better people can understand. So the more you explain about what policing means to a data scientists, the better the data scientist will understand what policing includes.”*

Expert 8 argued that: *“If you really want to understand how a decision is made, then it makes sense to look at the at the whole loop and of course explanations are then needed to understand what's happening.”* But when thinking about it more, she added that it is important to consider *“How do you design the explanations. Which parts are you going to highlight or emphasize or which information do you bring into the explanation, but also, what information do you leave out. When people explain things to one another then usually they don't give a full record of everything that is happening. They usually pick and choose those things that they think are important. If you make the wrong decisions in what you're going to explain or which information you're going to provide, then it can also be very confusing of course. Comprehensibility is lost. I mean designing a good explanation requires expertise.”*

Expert 8 was not the only one that doubted how evident the link was between giving explanations, even broad explanations, and improving the comprehensibility of predictive policing practice. Expert 4, for example, came back to her example of the data scientist and argued that: *“As a data scientist, you can try to somehow understand what policing is about, but if you've never been in the street and haven't encountered all the things that you can experience in a in a shift. You will never really understand what it means.”* Which led her to rephrase her initial claim that broad explainability could foster

comprehensibility: *“It will slightly foster it, but it will never really, you know, enhance it in a way that everybody will understand each other and.”*

Continuing this trend, Expert 2 argued that broad explainability could be fostering in some, but not in all cases, and Expert 1 argued that giving too much information in an explanation would not always improve comprehensibility as some people might not be able to process all that information. Which is also what Expert 9 concluded when explaining that: *“I think it's only comprehensible if you explain it in a way that it fits the audience you're targeting. I think it can be fostering, of course. But only if it fits the audience.”*

Only Expert 8 had chosen to discuss comprehensibility as an important value, but due to a technical failure of the recording equipment, the second part of the interview was lost. Lucky, the fill in document provided some notes from which some basic answers could be gathered. For audience, Expert 8 stated that explanations could be given to almost all people involved in the process of predictive policing, as well as people who are affected by potential crime and the judiciary. In terms of the related values, it has already been explained how some experts argued that comprehensibility relates to explainability and transparency. In the next section on trust we will also see that some experts argued that comprehensibility is related to trust. Finally, for the components of algorithmization, Expert 8 argued that all of them could be useful in explanations, dependent on whether a recipient could understand these components.

Relationship to Broad Explainability	Audience	Related Value	Components of Algorithmization
- Could be fostered as a better understanding of the specific context would allow for a better understanding of future explanations.	- People within the predictive policing process. - General Public. - Judiciary.	- Could benefit all values.	- All of them could be important.

4.2.2.6 Trust

Trust was defined as: *the extent to which people believe the predictive policing algorithm is treating them without prejudice and is working for their benefit.* When it came to trust, the experts first instinct was that broad explainability as a practice could foster this value. Expert 2 coined the term informed consent, explaining that even if someone does not fully support the practice or doesn't fully understand it. The fact that they are being communicated about how it works and how it's used can build trust in the

system and the people working with it. Several experts also emphasized that this is true for both people inside and people outside of the police organization. Expert 6 explained that, for police officers, if they are explained more about the context of a prediction then they would trust them more and be more inclined to follow them. Expert 4 pointed out that: *“If society would know how much people think about it before it gets used, then society would probably also trust it more because if it is thought through then it can be helpful instead of harmful.”*

However, when discussing this a bit further, many of the experts placed caveats with regards to how feasible this would actually be in practice with regards to citizen trust. Expert 4 stressed the point that it is very challenging in practice because you cannot really understand how trust comes about among citizens. Expert 5 argued that trust is only fostered if the explanations show that the practice is fair, otherwise it would have the opposite effect. Expert 7 made a similar argument, explaining that this is also dependent on the type of system the police is using (place- vs. person-based) which, according to his explanation, had to do with the extent to which a system could feel fair. Expert 8 argued that: *“Explanations in and of themselves don't increase trusts. And really, if the system itself is also correct and right and ethical, then it will increase trust. But of course, if the system is not trustworthy or unethical, then an explanation will hopefully decrease trusts and rightfully so.”* Which is in line with expert 9 who argued again that explanations should be adjusted based on audience or else they would not be effective in fostering values.

Seven experts were of the opinion that trust was an important value to foster in the context of predictive policing. The experts argued that trust could be fostered by giving both police officers and the general public explanations. Expert 1 argued that public trust had to be fostered even before the predictive policing system is used, and that this should be done by targeting explanations at everybody in the general public and not just those interested in the technology. Furthermore, Expert 2 argued that trust should also be fostered among the police, especially in light of the fact that a lot of police officers are reluctant to use the predictions of the system. Expert 5 summarized it best: *“Citizens because, if they distrust this technology, they will protest against the use of it and police officers because, if they don't trust it, they will not use it.”* Expert 10 also named politicians in power as important because they need to trust their police force to do their jobs right.

Pertaining to the related values, across the previously discussed values, it has already been shown that the experts argued that trust relates to effectiveness, accuracy, security, explainability, accountability, responsibility and transparency. Which is every value discussed up until now except efficiency and comprehensibility. But for both these values, a case was made when the values related to trust were discussed. Expert 2 made the argument that all the values relate to trust, which also included efficiency, comprehensibility and fairness: *“If you want to build trust, you need to be fair, you need to ensure privacy, you need to ensure comprehensibility, you need to ensure transparency and be accountable and responsible. Also as a second layer, also of course, it needs to be effective and accurate. So yeah, it's like cross cutting.”* Moreover, Expert 1 named security as important when fostering trust, and so did expert 5 who also named accuracy and privacy. Expert 6 and 9 also named privacy. With regards to these values, all of the experts named them as preconditional to trust in the sense that explanations need to show how these values are being considered in the predictive policing practice.

Finally, in terms of the components of algorithmization, arguments were made for all the components of algorithmization being relevant. However to varying degrees. Organizational structure, organizational policy, monitoring and evaluation as well as sociotechnical relations were most universally seen as important. Experts who took the standpoint of the general public often left out the technology and information relations components as they questioned whether these could be understood and they would not be fundamental for understanding how predictive policing works in practice. Although expert 9 did say that all of the components are important because trust is very complex and all of them are needed to understand the whole process and to fully be able to trust it. Experts taking the perspective from the police organization itself often did include technology and information relations as these are important to police officers. As Expert 2 argue: *“It's important for the officers who actually engage with technology, that they understand and feel it is part of the overall system that they work in and that it's not something outside their realm of control or sort of, their responsibilities. But it's an integral part of their daily organisation and the way police goes about its duties.”*

Relationship to Broad Explainability	Audience	Related Values	Components of Algorithmization
- Only the sense of Trust could be fostered. - Whether Trust is fostered is dependent on whether the	- Police officers. - General Public.	- Security, Effectiveness, Accuracy, and Privacy, could be helpful	- All of them could be important.

predictive policing process is trustworthy in the first place.			
--	--	--	--

4.2.2.7 *Fairness*

Fairness was defined as: *the extent to which a predictive policing algorithm considers everyone on the same basis*. In terms of fairness, the experts seemed split on whether broad explainability as a practice could foster this particular value. Experts 1, 4, 5, 6 all made similar arguments which could be summarized as follows: by giving explanations, people within the policing organization would have a better understanding of what is going on in the predictive policing process. Which would allow them to identify whether things are fair or not. Expert 4 gave the most elaborate explanation, stating that bias could enter the system in many ways. Historical training data could be tainted with bias if this data consists of old police reports made by police officers who were biased against a certain group. Similarly, the way predictions are being communicated could also be biased because in the end intelligence officers or other decision makers decide on how to allocate police resources based on risk predictions from the predictive policing algorithm, and they could be biased. Having broad explainability is beneficial because *“Then it becomes clear where the decisions are made about specific groups, who makes the decisions and at what point the decisions are made and why.”* Which would then allow the police to adjust practices that they identified as being unfair.

Expert 6 also argued that broad explanations could help to balance biases. She argued that, by explaining more about the context in of predictive policing, like the data selection process, decision makers could recognise it when, for example, only biased historical data is used. Expert 5 argued that broad explainability could make predictive policing practice more fair because, *“If you explain very well to the developers, for instance, that it should be fair, they can probably make it fair. Also, when you explain it to the citizens, they may feel it's used in a fair way.”* Similarly, Expert 1 argued that broad explanations might positively impact someone’s perception of how fair the predictive policing process is. But to this he did add that this is not guaranteed.

Experts 2, 3, 7, 8, and 9 were even less sure than Expert 1 on whether broad explainability could foster fairness. Expert 2 argued that explanations do not ensure non-discrimination and Expert 3 argued that only convincing explanations could foster fairness. Like Expert 1 and 6, Expert 7 also claimed that fairness is not necessarily an objective thing but rather a subjective thing. From this perspective expert 7 argued that

explanations might improve one's perception of how fair the process is, but that this would not change the fact that the process is discriminatory if it is. Which was also the argument by Expert 8: *"I can explain to you really accurately how I am biased against a certain group of people. And that doesn't make it fair."* Finally, expert 9 also argued that explanations might foster people's perception of fairness, but he argued that this was dependent on whether the process is fair in the first place.

Six experts (1, 2, 3, 4, 6, and 7) chose trust as one of the most important values to foster in the context of predictive policing. They argued that both people within the process of predictive policing as the general public could be targeted with explanations. Experts 1,2,3, and 6 emphasized the developers of the system as important because these would be able to implement fairness into the actual system. Besides this, also police officers were mentioned as important, and it would be the task of the developers to explain to them how fairness is incorporated in the system. Then Experts 4, 6, and 7 argued that the general public was also important to target with an explanation. In terms of the related values, it has already been explained that fairness relates to effectiveness, accuracy, accountability, transparency and trust. However, on the topic of how fairness relates to trust, Expert 9 had some additional comments. He explained that: *"I'm trusting someone because I think you make a fair decision or a fair judgement. So I think these are closely interrelated."* To which he later added: *"If the people don't experience fairness in the system. They automatically don't trust it, or at least, it makes it more likely for them to distrust the system."* Expert 9 also argued that explainability comes into play here. According to him: *"Giving explanations or explainability is a precondition for acceptance. Which necessitates fairness."*

Finally, links were also made with security and privacy. Expert 7 linked fairness to security and explained that there is a trade of relationship between the two. He explained that, as soon as you speak about the flexible allocation of resources in time and space then not everybody can get the same treatment. Which, he continued to explain, means that not everyone can be protected by predictive policing in the same sense and to the same extent. On the other hand, he added, showing that one neighbourhood actually needs to be more protected than others might convince people that there is a notion of fairness in the system. Expert 9 linked fairness to privacy and argued that if privacy is not properly taken into consideration then people could find the process unfair.

Lastly, with regards to the components of algorithmization, the most important components were technology, information relations and socio-technical relations. The

experts argued that these were most important to explain both internally and externally. To the police because these are the most important aspects of a process based on which one can judge whether something is done fair, and if not then this can be addressed. For the public this is the same, although these elements can only change their perception of fairness. Nevertheless, some experts argued again that all the components of algorithmization could be considered as important to be explained because fairness is something that pertains to the whole process.

Relationship to Broad Explainability	Audience	Related Value	Components of Algorithmization
<ul style="list-style-type: none"> - Could be fostered because the police might be able to improve the fairness of the process if they understand it better. - Could only foster the sense of fairness with citizens. Which is not guaranteed. 	<ul style="list-style-type: none"> - General public. - Police Officers. 	<ul style="list-style-type: none"> - Effectiveness, Accuracy, and Responsibility, could be helpful. 	<ul style="list-style-type: none"> - Technology, Information Relations, Socio-Technical Relations. - All of them could be important.

4.2.2.8 Privacy

Privacy was defined as: *the extent to which private data used for predictive policing is secure and untraceable*. According to Expert 3, the definition used in this thesis is very narrow and privacy should not only be about the security or traceability of private data that is being used but about data protection in general. Experts 1 and 4 also took an interesting perspective of privacy in the context of predictive policing. They also considered the privacy of people within the process of predictive policing. In the case of Expert 1 these are members of the judiciary and Expert 4 considered the privacy of police officers.

On the topic of whether broad explainability as a practice could foster privacy, the experts seemed to be split. Expert 1 argued that explanations would endanger privacy as these explanations could contain information that could be traced back to e.g. criminals or police officers. Expert 3 went the opposite direction, arguing that people have the right to get an account of how their data is used, which is an important part of data protection, and explanations are of course an important part of giving an account. Experts 4 and 9 gave descriptions of how explanations might foster privacy, but had to note that it could only potentially foster the experience of privacy. Both argued that by explaining about the entire process, people could get a better understanding of how their privacy is

protected in the process. Expert 4 argued from the perspective of the police and Expert 9 from the perspective of the general public. Expert 5 made a different argument, explaining that privacy is always breached by policing practice and that the real question is whether the breach of privacy is proportional to the security it provides. Broad explanations, she argued could help establish whether there is a proportional balance between privacy and security – a legitimate aim.

Four experts (3, 5, 6, and 9) chose privacy as an important value to foster. They again argued for both people within the predictive policing process as the general public as important audiences for explanations. Expert 5 was the only one who offered an elaborate explanation of how targeting these audiences could foster privacy. She explained that it is important to give explanations to citizens so that they could understand how their privacy might be violated and what they could do to protect their privacy. As for people within the predictive policing process, she argued that they would need explanations to understand how to establish a legitimate aim and in that sense proportionally balance privacy against security.

As for privacy’s relationship to other values, privacy was obviously named as related to security by expert 5 for reasons already explained. Besides this, it has already been explained how privacy relates to fairness, trust, transparency, responsibility, and accountability. Which are all the values with which privacy is linked, based on the responses of the experts. Lastly, all the components of algorithmization were named as important to include in explanations with arguments again pointing at the fact that to truly foster privacy through explanations, people need to get a comprehensive understanding of the predictive policing practice. Expert 5 put it best: *“It’s not only about the technology and how it works, but also about what you do with it. That depends on expertise, but also only information, relations, organisational structure, policies. I think it depends on all the components of algorithmization you distinguish.”*

Relationship to Broad Explainability	Audience	Related Value	Components of Algorithmization
- Could be fostering because the right to explanation is important for privacy. - Might also have to be balanced if explanations contain private information.	- People within the predictive policing process. - General public.	- Transparency, Accountability, and Responsibility, might be necessary. - Fairness and Trust, could be helpful.	- All of them could be important.

5 Discussion

This thesis aimed to fill the gap in our current knowledge on how explainability could foster values and, for that reason, conducted research on how explainability interacts with other values and how giving explanations could actually foster these values, in the context of predictive policing.

The main research question was:

How would experts describe the interaction between broad explainability and other important values in the context of predictive policing, and how, according to them, could explanations foster these values?

This thesis has specifically studied broad explainability - *the act of giving explanations that are aimed at explaining multiple aspects of the socio-technical context*. Which is relevant for the context of predictive policing as it has a higher potential of fostering the identified values than technology-centred explainability - *the act of giving explanations aimed at explaining the technology and how it makes decisions*. In the end, the conclusions drawn with regards to broad explainability, and how this value interacts with the identified values will also be applicable to other contexts of algorithm use. As such, the conclusions will be formulated in a general sense, rather than specifically tailored to the predictive policing context.

This section will discuss the results of the exploratory expert interviews and draw conclusions which will allow the researcher to develop an answer to the research question. First, this section will discuss how the experts perceived the interaction between broad explainability and the other values that were defined in the research background. After which, a discussion will be conducted for each of the factors of explanations – audience, related values, and, for content, the components of algorithmization. The conclusions drawn from these discussions will then be discussed in light of the research background in terms of their theoretical and practical implications. Finally, the limitations of this research will be discussed as well as some suggestions for future research.

5.1 Interpretation of the Results

5.1.1 Broad Explainability and Other Values: Fostering or Balancing?

In general all the experts believed that the values discussed in this thesis could be fostered by broad explainability. Only with effectiveness and efficiency were comments

made that this relationship might be balancing because the opportunity costs of explanations might be too high. As such, it is not relevant to further discuss the balancing relationship as no other relevant comments were made with regards to this. As for the fostering relationship, in many cases, the experts emphasized that they did not take this fostering relationship for granted

With regards to efficiency and effectiveness, the experts believed that giving explanations to the police could make them more efficient and effective, but stressed that there would be opportunity costs related to this. Furthermore, they also addressed the fact that if broad explanations are given to the public then this could cause criminals to adjust their behaviour in an attempt to cheat the system. Which means that in this instance, broad explanations should be balanced against efficiency and effectiveness.

For effectiveness, there were also some problems in terms of proving that the police is more effective. Which are also applicable to accuracy. In short, these problems address the difficulty in proving that the police is more effective or the algorithm more accurate. This is due to the fact that predictive policing is aimed at crime prevention, which means that to show that a predictive policing algorithm accurately predicted that a crime which was then prevented, one would have to provide proof of crimes that did not happen; which is inherently impossible. This could potentially explain why researchers have been having trouble attributing successful crime reductions to the use of predictive policing (Hunt, Saunders & Hollywood, 2014; Mohler et al., 2015; Levine, Tish, Tasso, and Joy, 2017). The expert did however think that broad explanations are given to the police then this could help them improve the accuracy of the predictive policing algorithm. Even though it is hard to prove.

Security was seen by the experts as the eventual goal of predictive policing, and they believed that security could be fostered. When arguing this, some of them claimed that security is not necessarily something that can be objectively measured and that broad explainability could only foster the feeling of security that people feel by making people trust the police more. Other experts argued that broad explainability could also foster security in more absolute terms. According to them, giving broad explanations to the police could actually help them better understand their practices which would also enable them to improve their practices and thus foster more security. Argued like this, the link between broad explainability and security lies in its potential to foster more effectiveness, efficiency and accuracy. Finally, the experts also argued that through broad

explainability, citizens would be enabled to take action with regards to ensuring their own safety. Which would also increase security.

Next is a group of values which are all very closely related: explainability (as a value), transparency and comprehensibility. For each of these values, the experts definitely thought that broad explainability could foster them, and this fostering relationship seemed to be the most direct of all the values. This is in line with a recent study conducted by Szczepański, Choraś, Pawlicki, and Pawlicka (2021), who also name interpretability, intelligibility, and understandability as concepts that closely relate to explainability.

Regarding explainability, the experts believed that broad explanations could foster the overall explainability of predictive policing as it would provide people more context and therefore they would be able to understand explanations better in the future. Furthermore, if (broad) explainability is set as a condition, it would also motivate the police to make their practices more explainable; this argument could be applied to every value. There is one condition to this however, and that is that the explanations must be relevant and comprehensible.

With regards to comprehensibility, the experts also believed that broad explainability would foster this value. They made similar arguments as with explainability, referring to the benefit of providing more context for the comprehensibility of future explanations. However, this fostering relationship is again subject to the same conditions as with explainability as a value: relevance and comprehensibility of the information. This is interesting as this would mean that comprehensibility is a condition for fostering comprehensibility. This circular reasoning can be broken if we phrase this relationship differently. If information that is shared is comprehensible, we can assume that it will lead to better understanding, which can then allow people to understand more complex and detailed information in the future – thus fostering comprehensibility. This shows that it is beneficial to think about giving explanations, not as something that you do once, but as something that can be done iteratively, which can eventually allow one to foster values better with explanations because understanding is improved with each iteration.

Finally, with regards to transparency, there was some ambiguity as to whether explainability is or isn't part of transparency. But on the basis of the discussion by the experts, it can be concluded that, in a theory, explainability should be a part of transparency because if something is not explainable then it can hardly be made

transparent. But in practice, we see that this is almost never the case and police organizations sometimes just incomprehensible data and call this transparency. Or they even use transparency as a tool to hide sensitive information. Which explains the emergence and need for explainability as a separate value, and broad explainability as a practice. Based on the close, almost intertwined, relationship between explainability and transparency, the experts believed that broad explainability would improve transparency.

Now we will discuss two other values that are closely related to each other, but also to the three previously mentioned values: accountability and responsibility. Again the experts believed that these values could be fostered by broad explainability, but, in the case of accountability, they again also argued that this would depend on whether information was comprehensible. Furthermore, with regards to accountability, they also argued that explaining the technology was not enough, because it might not be comprehensible, and that the police is accountable in the end so the whole practice should be scrutinised. Which are both arguments for broad explainability over technology-centred explainability. In terms of how explanations could foster accountability, the experts believed that giving explanations is an inherent part of accountability and that providing broader explanations could have a positive impact on accountability. On the other hand, they also believed that broad explanations given to the police themselves could help them understand better how to be accountable for their predictive policing practice.

Similar arguments were made with regards to how broad explainability could foster responsibility. The experts believed that broad explanations given to the police would allow them to better understand where their responsibilities lie. But even before that, broad explanations could make the police aware that they are responsible, e.g. for ensuring fairness and making sure that people trust the police's use of predictive policing will not harm them. Finally, they also believed that broad explanations could help the police identify breaches in responsibility, which could help them make their practices more responsible.

Then we have trust and fairness, another pair of values that are closely related to each other, but which also relate to all the previously mentioned values. As is the trend by now, the experts believed that trust and fairness could be fostered by broad explainability. However, looking at how the experts discussed these values, one can conclude that they are different from the previously discussed values in terms of how they are judged. With regards to the previously discussed values, the experts seem to generally

assume that these could somehow be measured or judged objectively. With the exception of security, none of the other values were discussed in terms of how they are perceived by people, i.e. whether the extent to which these values are present/realized depends on how people perceive this. But in the case of fairness and trust, the experts seem to think that, in terms of the general public, it is important to measure these values based on public perceptions of how fair and trustworthy the predictive policing practice is.

In terms of trust, the experts argued that that broad explainability could foster trust. They argued that giving explanations would foster trust because this could be the basis of informed consent with the public. However, they added to this that, when it comes to public trust, this is dependent on whether the police practice that is being explained is actually trustworthy and fair. If this is not the case than trust would not be fostered by broad explainability and it would actually be harmed. Which shows that public trust is judged based on how trustworthy and fair people perceive the predictive policing practice to be. Additionally, they also believed that broad explanations could foster trust within the police organisation itself, but in this case they didn't specify that it would only pertain to a feeling of trust. Instead they used the term acceptance.

The same is true for fairness, in which case the experts concluded that broad explainability might foster the perception of fairness but that this doesn't make the predictive policing practice more fair in absolute terms. However, on this point, some experts also argued that the fairness of predictive policing practice could actually be fostered in absolute terms, if broad explanations were given to the police themselves. It was argued that providing broad explanations to people within the predictive policing process themselves could allow them to spot instances in which the practice is not fair, which would allow them to make improve this and make it more fair.

Lastly, we have privacy, on the topic of whether broad explainability could foster privacy, the experts were split. Some of them believing that privacy could be harmed because private data could be shared with regards to people involved in the predictive policing practice. The others thinking that privacy could be fostered because explaining how private data is used is important part of privacy, if that includes data protection. One expert also argued that privacy is always breached and that broad explanations could foster privacy in the sense that they are needed to show that the breach of privacy was made on a legitimate basis. From this it can be argued that, similar to trust and fairness, privacy is mostly judged by people's perception of privacy rather than in actual objective terms.

Now based on this discussion of all the values and whether the results indicated if broad explainability could foster them, it can be concluded that all these values can potentially be fostered by broad explainability. However, the way in which they could be fostered and the directness with which they are fostered varies among these values. What is meant with 'directness with which they can be fostered' is easily explained. Some of the values – transparency, explainability and comprehensibility, as well as efficiency, effectiveness and accuracy – seemed to be thought of as being directly fostered by broad explainability. While for the other values – trust, fairness, security and privacy – there was some more doubt as to whether these values could truly be fostered by broad explainability.

Why this is the case can actually be explained by discussing the other factor along which these values vary: the way in which they are fostered. The group of values that are seen as potentially being directly fostered are also the ones that were not discussed as being dependent of people's perception of them. While for the other group of values it was always mentioned that, often in the case of the public, whether or not broad explainability would foster the value is dependent on what the public's perception of the value is.

This leads us to the first main finding of this thesis: *the extent to which one believes broad explainability can foster a certain value is dependent on whether the value is judged subjectively or objectively*. If a value is judged objectively, like efficiency, broad explanations could foster that value by improving people's, in this case the police's, understanding of the predictive policing process and allowing them to make the necessary changes to better realize that value. On the other hand, if a value is judged subjectively, like trust, then it is very hard to determine whether an explanation, even broad explanations, will foster someone's positive perception of that value. No matter how broad, or comprehensible, or well-designed an explanation is, there is no guarantee that it will improve someone's perception of a value, e.g. how trustworthy someone perceives the predictive policing process to be.

Now, it has to be added that all the values could be judged both objectively and subjectively. For example, we could objectively measure public trust by conducting a survey and having the public judge the trustworthiness of predictive policing on a Likert scale. Similarly, we could judge efficiency based on how efficient the public, or the police, perceive the predictive policing process to be. As such, as a general conclusion, one could say that broad explainability can foster values if they are judged objectively. If

they are judged subjectively then there is no guarantee and it could very well differ on a case by case basis. Finally, whether a value is judged objectively or subjectively is of course dependent on which person or organization is doing the judging; this will be discussed briefly in the section on Audience. For now it's important to take note of this distinction.

A second main finding can be derived from the previous discussion of all the values and their main results, which is that: *regardless of what kind of explainability (broad or technology-centred) or how the values are judged (subjective or objective) fostering values with explanations can only be done when some general conditions are met.* Three of these general conditions became apparent when discussing the values of explainability, comprehensibility and transparency. They are:

1. The information that is shared must be comprehensible.
2. The practice must be explainable.
3. The information that is shared must be relevant to fostering the value in question.

Now it might seem strange, that in a sense, both comprehensibility and explainability are both identified as a precondition for fostering values with broad explainability, as well as values that could be fostered by broad explainability. To address this, please remember that it was established earlier that it is beneficial to think about giving explanations, not as something that done once, but as something that can be done iteratively, which can eventually allow one to foster values better with explanations because understanding is improved with each iteration. Enabling people to understand more complex and detailed information in the future. As such, a basic level of comprehensibility and explainability can be set as a precondition for broad explainability, while broad explainability can foster higher levels of explainability and comprehensibility in the future over several iterations of explanation.

Analysing the comments that were made by the experts, relevant to our understanding of both broad and technology-centred explainability, a few more general conditions can be identified, which are:

4. The person or organization tasked with giving an explanation must be able and willing to share the information necessary to foster the value in question.
5. The person or organization tasked with giving the explanation must do so honestly.

6. The person or organization tasked with giving the explanation must actually have a mission to improve the value in question.

With regards to condition 4, one of the experts addressed the fact that sometimes the police is not willing to share certain information because it is of a sensitive nature. Sometimes because they are not sure whether what they are doing is legitimate or even the right way to do it and therefore they do not want that information to become public. This could be different if it was mandatory to give certain explanations, like the right to explanation in terms of data protection and privacy, as one of the experts mentioned. But for as far as the researcher knows, the right to broad explanations regarding predictive policing, or the public sectors use of algorithms has not been codified anywhere. Regarding the fifth condition, one of the experts mentioned that if you want to foster, e.g. trust, you have to give honest explanations. Even in the case that the predictive policing process is not as perfect as it should be, this could still foster trust because one is honest about it. Similar arguments can be made for the other values. Being dishonest about how fair, efficient, accurate, transparent etc. your predictive policing practice is, might foster the perception of these values. But this is a gamble that the truth is never found out, which in essence comes down to hiding malpractice, which in terms of public sector institutions is akin to suicide. Therefore, condition 4 can be seen as a necessary condition for truly fostering any of these values with (broad) explainability.

Finally, with regards to condition 5, the expert who was the largest inspiration of the previous two conditions also mentioned that the people giving the explanations must actually have a mission to accomplish certain objectives e.g. making the police more efficient or effective. This makes sense for all the values and also relates to the idea discussed in the research background that explanations must have a certain goal. Furthermore, this makes extra sense when considering that one of the experts explained that transparency is sometimes used as a tool to hide sensitive information about the predictive policing process – showing that fostering a value is definitely not always the goal of giving explanations.

5.1.2 How Broad Explanations can Foster Values: Factors of Explanation

5.1.2.1 Audience

In terms of audience, a separation can be made between two categories: (1) internal audience and (2) external audience. Internal audience refers to people who are

part of the predictive policing process and external audience refers to people who are outside the predictive policing process. Several different types of audiences were named during the interviews which are categorized in the following table:

Internal Audience	External Audience
Police officers	General public
Developers	Judiciary
Information officers	Defence lawyers
Decision makers (e.g. police top or politicians)	Journalists

There are two main findings with regards to the factor audience. The first one pertains to how these two audiences differ in terms of how they generally tend to judge values. Based on the arguments as to how broad explainability could foster different values, it can be concluded that: *internal audiences generally judge values based on objective standards, while external audiences more often judge them subjectively.* Consider the arguments that were made as to how broad explanations given to the police themselves would foster values like: efficiency, effectiveness, accuracy, but also transparency, accountability and responsibility, and even trust – although the trust of e.g. police officers who have to use the predictions was sometimes called acceptance. The arguments made could be summarized as: if they receive broad explanations, then this would make them understand the predictive policing practice better and that would enable them to make it more effective, accountable, transparent etc. Now consider the arguments made by the experts with regards to how broad explanations given to the public could foster values such as: trust, fairness, and security. The experts mostly argued that this was hard because it was hard to guarantee that an explanation would foster someone's positive perception of the value. Furthermore, they also argued that fostering someone's perception of a value does not mean that the value is fostered in an objective sense.

Based on this, it can be concluded that people within the predictive policing process might prefer to think about these values in objective terms. Which makes sense because they are responsible for the predictive policing practice and would be assessed based on how trustworthy transparent, fair etc. the practice is. Therefore, they would like to believe that they can impact the extent to which these values are realized and that they would be able to show this objectively. On the other hand, it can then also be concluded that, people outside the predictive policing process tend to think about these values in subjective terms. Which also makes sense because they have no direct influence on or

responsibility for the predictive policing practice. Therefore, they will judge the extent to which values are realized based on their own opinions. These can be well informed opinions, maybe based on some objective metrics showing how efficient or trustworthy etc. the police's use of predictive policing is, but it will still be subject to personal bias.

The second main finding with regards to audience builds on this, and has to do with the observation that these audiences, identified above, are not always the receivers of explanations. Sometimes they can also be the ones giving explanations. During the interviews, multiple times and by different experts, it was emphasized that developers were responsible for providing explanations on how the predictive policing algorithm works to the police officers. Furthermore, it was also mentioned that it would be beneficial for the developers to know more about police practice and what that involves. Then it was also said to be the responsibility of the police to provide explanations to the external audiences. This shows that in a sense, the responsibility for giving explanations is passed on throughout the different stages of the predictive policing process and, as such, the way in which we view the values (subjectively or objectively) also changes based on where we are in the process. Explanations from developers to the police will most likely focus on an objective view of the values, while explanations from the police to the external audiences might concern the values in a more subjective way. As such, we can explain the second main finding with regards to the audience as follows: *depending on where we are in the predictive policing process, the responsibility of giving explanations shifts and this also impacts whether the value is viewed objectively or subjectively.*

5.1.2.2 Related Values

During the second part of the interviews, the experts were asked to identify which values are related to a particular value, in the context of fostering that particular value. Furthermore, because all the values were discussed in the second part of the interview, the researcher was able to gather information pertaining to how all these values relate.

Four ways in which a value could relate to another value, in terms of fostering that particular value, were identified: (1) preconditional values, (2) helpful values, (3) potentially benefitted values, and (4) potentially endangered values. Preconditional values are values that need to be realized in order to foster this particular value. With regards to the preconditional values, based on the conditions for fostering values identified in this thesis, both explainability and comprehensibility can be considered to

always be preconditional to some extent for fostering all the values. This does not exclude them from being identified as having another type of relationship to other values as well, however it does mean that all values might potentially benefit from explainability and comprehensibility. Helpful values are values that could help foster a particular value but are not necessary. Potentially benefitted values are values that might also be fostered when a particular value is fostered. Lastly, potentially endangered values are values that might be endangered when a particular value is fostered. The values and their related values can be viewed in the table below:

Table 5.2: Important Values and their Related Values				
Value	Preconditional Values	Helpful Values	Potentially Benefitted Values	Potentially Endangered Values
<u>Efficiency</u>	Explainability, Comprehensibility	Accuracy, Accountability.	Effectiveness	
<u>Effectiveness</u>	Explainability, Comprehensibility	Efficiency, Accuracy	Trust, Responsibility, Accountability, Fairness	Transparency, Responsibility, Accountability
<u>Accuracy</u>	Explainability, Comprehensibility		Trust, Fairness, Efficiency, Effectiveness.	
<u>Security</u>	Explainability, Comprehensibility	Efficiency, Effectiveness, Accuracy	Trust	Privacy
<u>Explainability</u>	Explainability, Comprehensibility, Transparency		Trust, Responsibility, Transparency, Accountability, (All values)	
<u>Accountability</u>	Explainability, Comprehensibility, Transparency, Responsibility, Fairness, Accuracy		Transparency, Responsibility, Fairness, Trust	Effectiveness
<u>Responsibility</u>	Explainability, Comprehensibility, Transparency, Accountability		Accountability, Trust, Fairness	Effectiveness
<u>Transparency</u>	Explainability, Comprehensibility		Fairness, Trust, Privacy, Accountability, Explainability	Effectiveness
<u>Comprehensibility</u>	Explainability, Comprehensibility		(All values)	
<u>Trust</u>	Explainability, Comprehensibility, Transparency, Accountability, Responsibility	Security, Effectiveness, Accuracy, Privacy		Effectiveness
<u>Fairness</u>	Explainability, Comprehensibility, Accountability, Transparency	Effectiveness, Accuracy, Responsibility		
<u>Privacy</u>	Explainability, Comprehensibility, Transparency, Accountability, Responsibility	Fairness, Trust		

Now several caveats have to be placed with regards to this table of related values. Firstly, although every value was discussed in the second part of the interviews, not every value was picked by the same number of experts. Some values were picked by only one expert while others were picked by up to seven experts, meaning that those values that were picked more often were obviously discussed in much more depth than other values. Secondly, whether these values are preconditional, helpful, benefitted, or endangered may depend on whether the value is regarded objectively or subjectively. This can however not be determined based on the results of this thesis. Lastly, as can be seen in table 5.2 sometimes a value is placed in multiple categories with relation to the same value. For example transparency, responsibility and fairness were both identified as preconditional and potentially benefitted with regards to accountability. This means that the experts argued for multiple different types of relationships between these two values. What the conditions are for a value falling into one or the other category can also not be determined based on the results of this thesis.

Based on these three caveats, it must be concluded that the relationships between these values, that were identified from the results of this thesis, are incoherent and loosely substantiated. Which means they are highly hypothetical and that the relationships displayed above cannot be regarded as definite or proven and that more research on these relationships is needed. For these reasons, it is not deemed relevant to describe the different relationships between these values in detail. However, despite this, one more main finding can be determined with regards to how these values relate to each other: *values can be preconditional, helpful, potentially benefitted or potentially endangered when a certain value is fostered. This might be dependent on whether a value is viewed objectively or subjectively.*

5.1.2.3 Components of Algorithmization

Finally, with regards to the components of algorithmization, it must be said that for many of the values, the experts actually argued that all the components of algorithmization were important to explain. This could be due to the fact that the experts were asked to discuss the potential of broad explainability, which might have created a bias among the experts to perceive broader explanations, i.e. covering more aspects of the socio-technical context, as better/having a greater potential to foster values. There was however, some variation in the answers which will be shown in the table below. If some

of the experts argued that all the values could be important, then this is indicated by an ‘All’ between brackets.

Value	Components of Algorithmization
<u>Efficiency</u>	(All), Monitoring and Evaluation, Socio-Technical Relations and Expertise
<u>Effectiveness</u>	Technology, Organisational Structure, Socio-Technical Relations and Expertise
<u>Accuracy</u>	Technology
<u>Security</u>	(All)
<u>Explainability</u>	(All), Technology, Socio-Technical Relations, Information Relations
<u>Accountability</u>	(All)
<u>Responsibility</u>	(All)
<u>Transparency</u>	(All)
<u>Comprehensibility</u>	(All)
<u>Trust</u>	(All)
<u>Fairness</u>	(All), Technology, Information Relations, Socio-Technical Relations
<u>Privacy</u>	(All)

Based on the large number of ‘All’ displayed in the table above, it can be concluded that, according to the experts, explaining more is generally better. This is further supported by the fact that many of the experts specifically named Socio-Technical Relations, which are basically present in all stages and processes of predictive policing practice, as important. Even if they argued that all the components of algorithmization are important. However, this cannot be counted as a main finding because of the potential bias that the setup of the thesis might have caused.

Another argument that can be made with a little more certainty, is that there is a noticeable difference in the requirements for the content of explanations between internal and external audiences. When it came to internal audiences, the experts were generally of the opinion that it is important that they get a good understanding of how the predictive policing technology actually works. Besides which, they also believed it to be important that they get a good understanding of how they are supposed to work with the technology. In terms of external audiences, experts seemed to be more concerned with what values were demonstrated in the explanations, rather than which components of algorithmization were included. Because this build upon the distinction between internal and external audiences, it could also depend on whether values are judged on a subjective or objective basis. Which leads to the last main finding: *for internal explanations, it is relevant which*

components of algorithmization are included in an explanations, for external explanations it seems to be more relevant which values are demonstrated in an explanation. This might be dependent on whether a value is viewed objectively or subjectively.

5.2 Theoretical and Practical Implications

All in all, there were six main findings in this research:

1. The extent to which one believes that broad explainability can foster a certain value is dependent on whether the value is judged subjectively or objectively.
2. Regardless of what kind of explainability (broad or technology-centred) or how the values are judged (subjective or objective) fostering values with explanations can only be done when some general conditions are met.
3. Internal audiences generally judge values based on objective standards, while external audiences more often judge them subjectively.
4. Depending on where we are in the predictive policing process, the responsibility of giving explanations shifts and this also impacts whether the value is judged objectively or subjectively.
5. Values can be preconditional, helpful, potentially benefitted or potentially endangered when a certain value is fostered. This might be dependent on whether a value is viewed objectively or subjectively.
6. For internal explanations, it is relevant which components of algorithmization are included in an explanations, for external explanations it seems to be more relevant which values are demonstrated in an explanation. This might be dependent on whether a value is viewed objectively or subjectively.

The first two main findings has impact on our current understanding of (broad) explainability, and how, in a practical context, this could foster the values identified in this thesis. The remaining four provide new insights with regards to, and thus have impact on, the factors of explanation – audience, related values, and content – and how these impact the potential of (broad) explainability to foster the values identified in this thesis. In other words, they contribute to our understanding of what the right explanation could be to foster a certain value and what needs to be taken into account with regards to each of the factors of explanation. As such, this thesis has provided valuable contributions to the existing literature with regards to explainability and our current understanding of how

explainability interacts with a number of other values, and under what conditions these interactions could be fostering for those values.

In terms of their practical implications, the first main finding provides a new point of view for practitioners to consider when contemplating using (broad) explanations to foster certain values. The second main finding provides practitioners with a list of conditions that have to be met when wanting to foster values with (broad) explanations. The third main finding, provides a way to distinguish between different audiences and their particular view on values, which is relevant to consider if one wants to foster a certain value from their perspective. The fourth main finding shows that there is potentially a chain of explanations running throughout the process of using an algorithm which is relevant because it shows that how to foster values with (broad) explanations varies depending on where in the process of using predictive algorithms they are given. Which is something practitioners will have to take into account. The fifth main finding, outlines four different ways in which other values might relate to a particular value in the context of fostering that particular value. This is relevant for practice because, if an organization has certain core values, it is important to know how these are impacted when attempting to foster one of them with (broad) explanations. Finally, the sixth main finding, shows that there are different requirements for the content of explanations. Which might also be dependent on whether a value is viewed objectively or subjectively. This is relevant for practitioners because it helps them determine what the right content is for explanations fostering a certain value.

All in all, although this thesis has specifically focussed on the context of predictive policing. It should be mentioned again that the main findings of this thesis go beyond this specific context, as they are aimed at contributing to our general understanding of explainability. The researcher has taken special care to formulate the main findings in a way that they could be generally applied, also to other fields of (predictive) algorithm use in the public sector. The extent to which these findings are truly applicable beyond the context of predictive policing however, should be subject to future research and cannot be determined at this point. To conclude, the main findings from this thesis have made several valuable contributions to both the existing literature and existing practice with regards to explainability.

5.3 Limitations and Future Research

No research is free from limitations and as such, also this thesis is subject to a number of limitations. These limitations stem mostly from the methods and approaches that the researcher used in order to conduct research into the chosen topic. The first limitation that needs to be discussed is the abstractness of the research, combined with its practical focus. Several experts commented on the abstractness of the research and that it was a strenuous exercise to try and combine their knowledge of practice with the various concepts and perspectives that were defined in this thesis. Although none of the experts commented that it was too hard or that they were unable to answer the interview question, most of them did need additional explanation during the interviews which further limited the already limited time. In the end, this did not prevent the researcher from acquiring a solid amount of relevant data, but it did mean that some of the topics could not be discussed in as much depth as would be preferable.

This leads to a second limitation which also impacted how in depth all the topics could be discussed during the interviews: the scope of the research. The scope of this thesis is quite broad, having defined a new form of explainability, 12 important values, two ways in which (broad) explainability could relate to these values, 3 relevant factors of explanation and 7 components of algorithmization. All in all, there was a lot to discuss in the interviews. The large number of topics to cover in the interviews was remedied by a well organised interview set up and a preparatory document which enabled the experts to familiarize themselves with the research background before the interviews.

As was mentioned in the methodology, a number of limitation stems from the chosen research method: exploratory expert interviews. The first one is that this methodology means that the insights that were derived from this study are largely hypothetical and sometimes a bit shallow as there was not a lot of room to discuss specific practical examples in depth. On the other hand, this method did allow for a broad discussion of the different topics and, as such, a broad picture of (broad) explainability and its interactions with other values in the context of predictive policing. Which eventually led to the formulation of the six main findings which, due to this broad basis, have implications beyond the specific context of predictive policing.

Some limitations also stem from the expert sample that was selected. The first limitation emerges from the sample characteristics. Experts from a broad range of different research backgrounds were chosen to participate in this study. Which again

widens the scope of the research, which causes the main findings to be a bit under-substantiated. However this was necessary to paint the broad picture of explainability in the context of predictive policing that was needed to produce the main findings which are broadly applicable to different contexts of algorithm use by the public sector. Another limitation from the expert sample stems from its size. Preferably, the sample size would have been larger than the 10 experts, as this would allow for even more in depth study of the subject. However, the sample size of this thesis is large enough to provide the necessary basis for studying the subject of this thesis and to provide a solid basis for some interesting findings that, in turn, provide a good basis for future research.

As for suggestion for future research, several can be made, based on the main findings of this thesis. The first suggestion for future research is to look more into the impact of viewing values subjectively or objectively on the potential of (broad) explainability to foster values. Secondly, more research can be conducted on the conditions for (broad) explainability to foster different values in a practical context, and their applicability to different contexts of algorithm use by the public sector. Thirdly, more research can also be conducted on the differences between audiences and what this means in terms of fostering values with (broad) explanations ; related to the subjective objective distinction with regards to value. Fourthly, more research can be conducted on the so called ‘chains of explanations’ between different actors and at different stages of the use of algorithms by the public sector. As well as how these manifest in different instances of algorithm use by the public sector. Then finally, more research is definitely needed on how different values relate to each other when fostering a particular value in a practical context of algorithm use by the public sector, as well as on the different requirements with regards to the content of explanations, aiming to foster certain values, and how this varies depending on the other factors of explanation.

Besides these finding specific recommendations, a few general recommendations can also be made. More research should be done on the exact meaning of different values in different contexts of algorithm use by the public sector. Furthermore, mor research should be done on how variations in the underlying predictive models impacts the potential of realising (broad) explainability. Which leads to the final suggestion for future research, which is to, in general, conduct more case studies of the use of explanations for fostering values, in different contexts of algorithm use by the public sector.

Conclusion

This thesis set out to address the gap in our current knowledge on how explainability could foster values in a practical context, and to conduct research on how explainability interacts with other values and how giving explanations could actually foster these values in a practical context. For this purpose, the following research question was defined:

How would experts describe the interaction between broad explainability and other important values in the context of predictive policing, and how, according to them, could explanations foster these values?

In order to enable research on this topic, a solid theoretical fundament needed to be built in the research background. First, 12 values were identified, after which, the concept of broad explainability – *the act of giving explanations that are aimed at explaining multiple aspects of the socio-technical context* – was defined. Additionally, two ways in which broad explainability could relate to predictive policing were outlined: fostering and balancing, as well as, three factors of explanation – audience, related values, and content. Finally, on the topic of content, 7 components of algorithmization along, which the content of broad explanations could be structured, were also described.

All of these elements formed the basis of the exploratory expert interviews that were conducted in order to gather the data needed to develop an answer to the main research question. These exploratory expert interviews consisted of two parts. In the first part, the experts were questioned on whether they believed the relationship between the 12 identified values and broad explainability is fostering and when this relationship is more balancing. In the second part, they were asked to pick the five values that they believed were most important to foster in the context of predictive policing. After which they were questioned on the factors of explanation and what these would entail in the context of fostering each of the 5 chosen values. All the interviews were recorded and subsequently transcribed. The transcriptions were then analysed and interpreted with the aid of a coding software: MAXQDA.

Based on the interpretation of the results, 6 main findings were described:

1. The extent to which one believes that broad explainability can foster a certain value is dependent on whether the value is judged subjectively or objectively.

2. Regardless of what kind of explainability (broad or technology-centred) or how the values are judged (subjective or objective) fostering values with explanations can only be done when some general conditions are met.
3. Internal audiences generally judge values based on objective standards, while external audiences more often judge them subjectively.
4. Depending on where we are in the predictive policing process, the responsibility of giving explanations shifts and this also impacts whether the value is judged objectively or subjectively.
5. Values can be preconditional, helpful, potentially benefitted or potentially endangered when a certain value is fostered. This might be dependent on whether a value is viewed objectively or subjectively.
6. For internal explanations, it is relevant which components of algorithmization are included in an explanations, for external explanations it seems to be more relevant which values are demonstrated in an explanation. This might be dependent on whether a value is viewed objectively or subjectively.

The main findings from this thesis have made several valuable contributions to both the existing literature and existing practice with regards to explainability. The first two findings providing an answer to the first part of the research question, and giving insights with regards to how (broad) explainability related to other values. The remaining four findings providing an answer to the second part of the research question, and providing insights with regards to the factors of explanation – audience, related values, and content – and how those determine what type of explanations could foster certain values.

Finally, despite a number of significant limitation which were outlined, the findings of this research can provide a solid basis for future research on how explainability in different practical contexts. It provides a large number of new leads and approaches that can be researched and tested, which will continue to expand on our knowledge of explainability as a value, and the potential of (broad) explainability to foster different values. An effort that is not unimportant, because our current conception of mainly techno-centred explainability methods are unsuitable to deliver on the great promise of explainability and the hopes that practitioners and public sector organisations have placed in it.

References

- Alikhademi, K., Drobina, E., Prioleau, D., Richardson, B., Purves, D., & Gilbert, J. E. (2021). A review of predictive policing from the perspective of fairness. *Artificial Intelligence and Law*. <https://doi.org/10.1007/s10506-021-09286-4>
- Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2477899>
- Bennet Moses, L., & Chan, J. (2018). Algorithmic prediction in policing: Assumptions, evaluation, and accountability. *Policing and Society*, 28(7), 806–822. <https://doi.org/10.1080/10439463.2016.1253695>
- Bertossi, L., & Geerts, F. (2020). Data Quality and Explainable AI. *Journal of Data and Information Quality*, 12(2), 1–9.
- Bogner, A., & Menz, W. (2009). The Theory-Generating Expert Interview: Epistemological Interest, Forms of Knowledge, Interaction. In A. Bogner, B. Littig, & W. Menz (Eds.), *Interviewing Experts* (pp. 43–80). Palgrave Macmillan UK. https://doi.org/10.1057/9780230244276_3
- Bovens, M., Goodin, R. E., Schillemans, T., Bovens, M., Schillemans, T., & Goodin, R. E. (2014). Public Accountability. In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), *The Oxford Handbook of Public Accountability*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780199641253.013.0012>
- Brayne, S. (2017). Big Data Surveillance: The Case of Policing. *American Sociological Review*, 911–1008.
- Chan, J., & Bennett Moses, L. (2016). Is Big Data challenging criminology? *Theoretical Criminology*, 20(1), 21–39. <https://doi.org/10.1177/1362480615586614>
- Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda.

International Journal of Information Management, 60, 102383.

<https://doi.org/10.1016/j.ijinfomgt.2021.102383>

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic Decision Making and the Cost of Fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806.

<https://doi.org/10.1145/3097983.3098095>

de Laat, P. B. (2019). The disciplinary power of predictive algorithms: A Foucauldian perspective. *Ethics and Information Technology*, 21(4), 319–329.

<https://doi.org/10.1007/s10676-019-09509-y>

Eubanks, V. (2017). *Automating inequality: How high-tech tools profile, police, and punish the poor* (First Edition). St. Martin's Press.

Ferguson, A. G. (2017). *The rise of big data policing: Surveillance, race, and the future of law enforcement*. New York University Press.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining Explanations: An Overview of Interpretability of Machine Learning. *The 5th IEEE International Conference on Data Science and Advanced Analytics*.

<http://arxiv.org/abs/1806.00069>

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30(1), 99–120. <https://doi.org/10.1007/s11023-020-09517-8>

Hälterlein, J. (2021). Epistemologies of predictive policing: Mathematical social science, social physics and machine learning. *Big Data & Society*, 8(1), 205395172110031.

<https://doi.org/10.1177/20539517211003118>

Hansen, L. K., & Rieger, L. (2019). Interpretability in Intelligent Systems – A New Concept? In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.),

Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (pp. 41–49). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_3

Haque, M. R., Weathington, K., Chudzik, J., & Guha, S. (2020). Understanding Law Enforcement and Common Peoples' Perspectives on Designing Explainable Crime Mapping Algorithms. *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, 269–273. <https://doi.org/10.1145/3406865.3418330>

Hardyns, W., & Rummens, A. (2018). Predictive Policing as a New Tool for Law Enforcement? Recent Developments and Challenges. *European Journal on Criminal Policy and Research*, 24(3), 201–218. <https://doi.org/10.1007/s10610-017-9361-2>

Henriksen, A., & Bechmann, A. (2020). Building truths in AI: Making predictive algorithms doable in healthcare. *Information, Communication & Society*, 23(6), 802–816. <https://doi.org/10.1080/1369118X.2020.1751866>

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7), 900–915. <https://doi.org/10.1080/1369118X.2019.1573912>

Hunt, P., Saunders, J., & Hollywood, J. S. (2014). Conclusions. In *Evaluation of the Shreveport Predictive Policing Experiment* (pp. 49–52). RAND Corporation. <https://www.jstor.org/stable/10.7249/j.ctt14bs27t.12>

Janssen, M., & van den Hoven, J. (2015). Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy? *Government Information Quarterly*, 32(4), 363–368. <https://doi.org/10.1016/j.giq.2015.11.007>

Kasapoglu, T., & Masso, A. (2021). Attaining Security Through Algorithms: Perspectives of Refugees and Data Experts. In J. B. Wiest (Ed.), *Studies in Media and Communications*

(pp. 47–65). Emerald Publishing Limited. <https://doi.org/10.1108/S2050-206020210000020009>

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human Decisions and Machine Predictions*. *The Quarterly Journal of Economics*.

<https://doi.org/10.1093/qje/qjx032>

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy & Technology*, 31(4), 611–627.

<https://doi.org/10.1007/s13347-017-0279-x>

Levine, E. S., Tisch, J., Tasso, A., & Joy, M. (2017). The New York City Police Department's Domain Awareness System. *INFORMS Journal on Applied Analytics*, 47(1), 70–84. <https://doi.org/10.1287/inte.2016.0860>

Littig, B. (2009). Interviewing the Elite — Interviewing Experts: Is There a Difference? In A. Bogner, B. Littig, & W. Menz (Eds.), *Interviewing Experts* (pp. 98–113). Palgrave Macmillan UK. https://doi.org/10.1057/9780230244276_5

Maguire, M. (2014). Socio-technical systems and interaction design – 21st century relevance. *Applied Ergonomics*, 45(2), 162–170. <https://doi.org/10.1016/j.apergo.2013.05.011>

McCarthy, O. J. (2019). *AI & Global Governance: Turning the Tide on Crime with Predictive Policing* - United Nations University Centre for Policy Research. <https://cpr.unu.edu/publications/articles/ai-global-governance-turning-the-tide-on-crime-with-predictive-policing.html>

Meijer, A., & Grimmelikhuijsen, S. (2020). Responsible and accountable algorithmization: How to generate citizen trust in governmental usage of algorithms. In *The Algorithmic Society*. Routledge.

- Meijer, A., & Wessels, M. (2019). Predictive Policing: Review of Benefits and Drawbacks. *International Journal of Public Administration*, 42(12), 1031–1039.
<https://doi.org/10.1080/01900692.2019.1575664>
- Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation* (Fourth edition). John Wiley & Sons.
- Meuser, M., & Nagel, U. (2009). The Expert Interview and Changes in Knowledge Production. In A. Bogner, B. Littig, & W. Menz (Eds.), *Interviewing Experts* (pp. 17–42). Palgrave Macmillan UK. https://doi.org/10.1057/9780230244276_2
- Miailhe, N., & Hodes, C. (2017). *Making the AI revolution work for everyone* (p. 29). The Future Society at Harvard Kennedy School of Government. <http://ai-initiative.org/wp-content/uploads/2017/08/Making-the-AI-Revolution-workfor-everyone.-Report-to-OECD.-MARCH-2017.pdf>.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J. (2015). Randomized Controlled Field Trials of Predictive Policing. *Journal of the American Statistical Association*, 110(512), 1399–1411.
<https://doi.org/10.1080/01621459.2015.1077710>
- OECD. (2019). *Hello, World: Artificial intelligence and its use in the public sector* (OECD Working Papers on Public Governance No. 36; OECD Working Papers on Public Governance, Vol. 36). <https://doi.org/10.1787/726fd39d-en>
- Ogunleye, J. (2014). The Concepts of Predictive Analytics. *International Journal of Knowledge, Innovation, and Entrepreneurship*, 2(2), 82–90.
- O’Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy* (First edition). Crown.

- Paszynski, M., Kranzlmüller, D., Krzhizhanovskaya, V. V., Dongarra, J. J., & Sloot, P. M. A. (Eds.). (2021). *Computational Science – ICCS 2021: 21st International Conference, Krakow, Poland, June 16–18, 2021, Proceedings, Part IV* (Vol. 12745). Springer International Publishing. <https://doi.org/10.1007/978-3-030-77970-2>
- Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice* (Fourth edition). SAGE Publications, Inc.
- Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), 137–141. <https://doi.org/10.1007/s11747-019-00710-5>
- Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can You Recognize an Effective Teacher When You Recruit One? *Education Finance and Policy*, 6(1), 43–74.
- Russell, S. J., Norvig, P., & Davis, E. (2016). *Artificial intelligence: A modern approach* (3rd ed). Prentice Hall.
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2020). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. SPRINGER NATURE.
- Samek, W., & Müller, K.-R. (2019). Towards Explainable Artificial Intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 5–22). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_1
- Sandhu, A., & Fussey, P. (2021). The ‘uberization of policing’? How police negotiate and operationalise predictive policing technology. *Policing and Society*, 31(1), 66–81. <https://doi.org/10.1080/10439463.2020.1803315>
- Santos, R. B. (2014). The Effectiveness of Crime Analysis for Crime Reduction: Cure or Diagnosis? *Journal of Contemporary Criminal Justice*, 30(2), 147–168. <https://doi.org/10.1177/1043986214525080>

- Shaban-Nejad, A., Michalowski, M., & Buckeridge, D. L. (2021). Explainability and Interpretability: Keys to Deep Medicine. In A. Shaban-Nejad, M. Michalowski, & D. L. Buckeridge (Eds.), *Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability* (pp. 1–10). Springer International Publishing. https://doi.org/10.1007/978-3-030-53352-6_1
- Szczepański, M., Choraś, M., Pawlicki, M., & Pawlicka, A. (2021). The Methods and Approaches of Explainable Artificial Intelligence. In M. Paszynski, D. Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, & P. M. A. Sloot (Eds.), *Computational Science – ICCS 2021* (pp. 3–17). Springer International Publishing. https://doi.org/10.1007/978-3-030-77970-2_1
- Tene, O., & Polonetsky, J. (2017). TAMING THE GOLEM: CHALLENGES OF ETHICAL ALGORITHMIC DECISION-MAKING. *North Carolina Journal of Law & Technology*, 19(1), 125–vi.
- Thierer, A., O’Sullivan, A. C., & Russell, R. (2017). *Artificial Intelligence and Public Policy* (p. 56) [Study]. Mercatus Centre. <https://www.mercatus.org/publications/artificial-intelligence-public-policy>
- van Brakel, R. (2016). Pre-Emptive Big Data Surveillance and its (Dis)Empowering Consequences: The Case of Predictive Policing. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2772469>
- Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. <https://doi.org/10.1145/3173574.3174014>
- Wartell, J., & McEwen, J. T. (2001). *Privacy in the Information Age: A Guide for Sharing Crime Maps and Spatial data*. National Insitute of Justice.

- Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a literature Review. *MIS Quarterly*, 26(2), xiii–xxiii.
- Weller, A. (2019). Transparency: Motivations and Challenges. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 23–40). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_2
- Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 1–18.
<https://doi.org/10.1145/3351095.3372833>
- Zarsky, T. (2016). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science, Technology, & Human Values*, 41(1), 118–132.
<https://doi.org/10.1177/0162243915605575>
- Zuiderveen Borgesius, F. (2018). *Discrimination, artificial intelligence, and algorithmic decision-making* (p. 52) [Study]. Council of Europe.

Appendix

A Concept Matrix

Clarification of the Concept Labels

Label	Concept
<u>A</u>	Explainability
<u>B</u>	AI
<u>C</u>	Explainable AI
<u>D</u>	Algorithmic Decision Making
<u>E</u>	Public Sector
<u>F</u>	Predictive Algorithms
<u>G</u>	Predictive Policing
<u>H</u>	Fairness
<u>I</u>	Discrimination
<u>J</u>	Accountability
<u>K</u>	Transparency

Article	Applicable concepts										
	A	B	C	D	E	F	G	H	I	J	K
Alikhademi, K., Drobinina, E., Prioleau, D., Richardson, B., Purves, D., & Gilbert, J. E. (2021). A review of predictive policing from the perspective of fairness. <i>Artificial Intelligence and Law</i> . https://doi.org/10.1007/s10506-021-09286-4		X					X	X			
Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2477899				X		X			X		
Bennet Moses, L., & Chan, J. (2018). Algorithmic prediction in policing: Assumptions, evaluation, and accountability. <i>Policing and Society</i> , 28(7), 806–822. https://doi.org/10.1080/10439463.2016.1253695						X	X			X	
Bertossi, L., & Geerts, F. (2020). Data	X	X							X		

Quality and Explainable AI. Journal of Data and Information Quality, 12(2), 1–9.											
Bovens, M., Goodin, R. E., Schillemans, T., Bovens, M., Schillemans, T., & Goodin, R. E. (2014). Public Accountability. In M. Bovens, R. E. Goodin, & T. Schillemans (Eds.), The Oxford Handbook of Public Accountability. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199641253.013.0012				X						X	
Brayne, S. (2017). Big Data Surveillance: The Case of Policing. American Sociological Review, 911–1008.					x		X				
Chan, J., & Bennett Moses, L. (2016). Is Big Data challenging criminology? Theoretical Criminology, 20(1), 21–39. https://doi.org/10.1177/1362480615586614		X				X	X				

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. <i>International Journal of Information Management</i> , 60, 102383. https://doi.org/10.1016/j.ijinfomgt.2021.102383		X									
Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic Decision Making and the Cost of Fairness. <i>Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining</i> , 797–806. https://doi.org/10.1145/3097983.3098095		X		X				X	X		
de Laat, P. B. (2019). The disciplinary power of predictive algorithms: A Foucauldian perspective. <i>Ethics</i>		X				X					X

and Information Technology, 21(4), 319–329. https://doi.org/10.1007/s10676-019-09509-y											
Eubanks, V. (2017). Automating inequality: How high-tech tools profile, police, and punish the poor (First Edition). St. Martin's Press.				X		X	X	X	X	X	X
Ferguson, A. G. (2017). The rise of big data policing: Surveillance, race, and the future of law enforcement. New York University Press.		X				X	X	X	X	X	X
Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining Explanations: An Overview of Interpretability of Machine Learning. The 5th IEEE International Conference on Data Science and Advanced Analytics. http://arxiv.org/abs/1806.00069	X	X	X								

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. <i>Minds and Machines</i> , 30(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8		X	X					X	X	X	X
Hälterlein, J. (2021). Epistemologies of predictive policing: Mathematical social science, social physics and machine learning. <i>Big Data & Society</i> , 8(1). 205395172110031. https://doi.org/10.1177/20539517211003118						X	X				
Hansen, L. K., & Rieger, L. (2019). Interpretability in Intelligent Systems – A New Concept? In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), <i>Explainable AI: Interpreting, Explaining and Visualizing Deep</i>	X		X								

Learning (pp. 41–49). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_3											
Haque, M. R., Weathington, K., Chudzik, J., & Guha, S. (2020). Understanding Law Enforcement and Common Peoples' Perspectives on Designing Explainable Crime Mapping Algorithms. Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing, 269–273. https://doi.org/10.1145/3406865.3418330	X		X				X				
Hardyns, W., & Rummens, A. (2018). Predictive Policing as a New Tool for Law Enforcement? Recent Developments and Challenges. European Journal on Criminal Policy and Research, 24(3), 201–218.							X				

https://doi.org/10.1007/s10610-017-9361-2											
Henriksen, A., & Bechmann, A. (2020). Building truths in AI: Making predictive algorithms doable in healthcare. <i>Information, Communication & Society</i> , 23(6), 802–816. https://doi.org/10.1080/1369118X.2020.1751866	X	X		X							
Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. <i>Information, Communication & Society</i> , 22(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912		X						X	X		
Hunt, P., Saunders, J., & Hollywood, J. S. (2014). Conclusions. In <i>Evaluation of the Shreveport Predictive Policing Experiment</i> (pp. 49–52). RAND Corporation.							X				

https://www.jstor.org/stable/10.7249/j.ctt14bs27t.12											
Janssen, M., & van den Hoven, J. (2015). Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy? <i>Government Information Quarterly</i> , 32(4), 363–368. https://doi.org/10.1016/j.giq.2015.11.007					X						X
Kasapoglu, T., & Masso, A. (2021). Attaining Security Through Algorithms: Perspectives of Refugees and Data Experts. In J. B. Wiest (Ed.), <i>Studies in Media and Communications</i> (pp. 47–65). Emerald Publishing Limited. https://doi.org/10.1108/S2050-206020210000020009							X	X			
Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human Decisions and				X		X					

<p>21st century relevance. Applied Ergonomics, 45(2), 162–170. https://doi.org/10.1016/j.apergo.2013.05.011</p>											
<p>McCarthy, O. J. (2019). AI & Global Governance: Turning the Tide on Crime with Predictive Policing - United Nations University Centre for Policy Research. https://cpr.unu.edu/publications/articles/ai-global-governance-turning-the-tide-on-crime-with-predictive-policing.html</p>		X					X				
<p>Meijer, A., & Grimmelikhuijsen, S. (2020). Responsible and accountable algorithmization: How to generate citizen trust in governmental usage of algorithms. In <i>The Algorithmic Society</i>. Routledge.</p>		X						X		X	
<p>Meijer, A., & Wessels, M. (2019). Predictive Policing: Review of Benefits</p>							X				

and Drawbacks. International Journal of Public Administration, 42(12), 1031–1039. https://doi.org/10.1080/01900692.2019.1575664											
Miailhe, N., & Hodes, C. (2017). Making the AI revolution work for everyone (p. 29). The Future Society at Harvard Kennedy School of Government. http://ai-initiative.org/wp-content/uploads/2017/08/Making-the-AI-Revolution-work-for-everyone.-Report-to-OECD.-MARCH-2017.pdf .		X									
Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. <i>Artificial Intelligence</i> , 267, 1–38. https://doi.org/10.1016/j.artint.2018.07.007	X	X	X								
Mohler, G. O., Short, M. B., Malinowski, S., Johnson, M., Tita, G. E., Bertozzi, A. L., & Brantingham, P. J.							X				

(2015). Randomized Controlled Field Trials of Predictive Policing. Journal of the American Statistical Association, 110(512), 1399–1411. https://doi.org/10.1080/01621459.2015.1077710											
O’Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy (First edition). Crown.		X		X				X		X	X
OECD. (2019). Hello, World: Artificial intelligence and its use in the public sector (OECD Working Papers on Public Governance No. 36; OECD Working Papers on Public Governance, Vol. 36). https://doi.org/10.1787/726fd39d-en		X									
Ogunleye, J. (2014). The Concepts of Predictive Analytics. International Journal of Knowledge, Innovation, and						X					

Entrepreneurship, 2(2), 82–90.											
Rai, A. (2020). Explainable AI: From black box to glass box. <i>Journal of the Academy of Marketing Science</i> , 48(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5	X	X	X								
Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can You Recognize an Effective Teacher When You Recruit One? <i>Education Finance and Policy</i> , 6(1), 43–74.				X	X	X					
Russell, S. J., Norvig, P., & Davis, E. (2016). <i>Artificial intelligence: A modern approach</i> (3rd ed). Prentice Hall.		X									
Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2020). <i>Explainable AI: Interpreting, Explaining and Visualizing Deep</i>	X	X	X	X				X		X	X

Learning. SPRINGER NATURE.											
Samek, W., & Müller, K.-R. (2019). Towards Explainable Artificial Intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (pp. 5–22). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_1	X	x	X					X		X	X
Sandhu, A., & Fussey, P. (2021). The ‘uberization of policing’? How police negotiate and operationalise predictive policing technology. Policing and Society, 31(1), 66–81. https://doi.org/10.1080/10439463.2020.1803315	X						X				
Santos, R. B. (2014). The Effectiveness of Crime Analysis for							X				

Crime Reduction: Cure or Diagnosis? Journal of Contemporary Criminal Justice, 30(2), 147–168. https://doi.org/10.1177/1043986214525080											
Shaban-Nejad, A., Michalowski, M., & Buckeridge, D. L. (2021). Explainability and Interpretability: Keys to Deep Medicine. In A. Shaban-Nejad, M. Michalowski, & D. L. Buckeridge (Eds.), Explainable AI in Healthcare and Medicine: Building a Culture of Transparency and Accountability (pp. 1– 10). Springer International Publishing. https://doi.org/10.1007/978-3-030-53352-6_1	X	X	X								
Szczeptański, M., Choraś, M., Pawlicki, M., & Pawlicka, A. (2021). The Methods and Approaches of Explainable Artificial Intelligence. In M. Paszynski, D.	X	X	X	X							

<p>Kranzlmüller, V. V. Krzhizhanovskaya, J. J. Dongarra, & P. M. A. Sloot (Eds.), Computational Science – ICCS 2021 (pp. 3–17). Springer International Publishing. https://doi.org/10.1007/978-3-030-77970-2_1</p>											
<p>Tene, O., & Polonetsky, J. (2017). TAMING THE GOLEM: CHALLENGES OF ETHICAL ALGORITHMIC DECISION-MAKING. North Carolina Journal of Law & Technology, 19(1), 125–vi.</p>		X	X					X	X	X	X
<p>Thierer, A., O’Sullivan, A. C., & Russell, R. (2017). Artificial Intelligence and Public Policy (p. 56) [Study]. Mercatus Centre. https://www.mercatus.org/publications/artificial-intelligence-public-policy</p>		X									
<p>van Brakel, R. (2016). Pre-Emptive Big Data</p>							X				

Surveillance and its (Dis)Empowering Consequences: The Case of Predictive Policing. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2772469											
Veale, M., Van Kleek, M., & Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 1–14. https://doi.org/10.1145/3173574.3174014		X		X				X	X	X	X
Wartell, J., & McEwen, J. T. (2001). Privacy in the Information Age: A Guide for Sharing Crime Maps and Spatial data. National Insitute of Justice.		X									
Weller, A. (2019). Transparency: Motivations and Challenges. In W. Samek, G. Montavon, A. Vedaldi, L. K.	X	X	X	X							X

Hansen, & K.-R. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (pp. 23–40). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_2											
Wieringa, M. (2020). What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 1–18. https://doi.org/10.1145/3351095.3372833		X								X	X
Zarsky, T. (2016). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. Science, Technology, & Human Values, 41(1),		X		X				X	X		X

118–132. https://doi.org/10.1177/0162243915605575											
Zuiderveen Borgesius, F. (2018). Discrimination, artificial intelligence, and algorithmic decision-making (p. 52) [Study]. Council of Europe.		X		X				X	X	X	X

B Interview Schedule

Part 1: General Information	
Focus Area	Example questions
<u>Expertise of the interviewee</u>	<ul style="list-style-type: none"> - Please tell me about yourself and your background. - Please tell me from which point of view you have engaged with the topic of predictive policing.
<u>Interview preparation</u>	<ul style="list-style-type: none"> - Was there anything in the interview preparation which was unclear and in need of further elaboration?
Part 2: Main Values at stake and their relationship with explainability	
<u>Definition and relevance of the identified values</u>	<ul style="list-style-type: none"> - Could you please tell me whether you agree with the definitions of the values given to you in the preparation?
<u>Relationship</u>	<ul style="list-style-type: none"> - Could you please indicate for each of these values whether they have a more balancing or fostering relationship to explainability?
<u>Importance</u>	<ul style="list-style-type: none"> - Which of these values do you believe are most important to foster in the context of predictive policing practice? (Top 5) And why?
Part 3: Determining which explanations can foster which (Top 5) value	
<u>Intended Audience</u>	<ul style="list-style-type: none"> - For each of these values, could you please indicate who the intended audience of explainability could be if one wants to foster that particular value?
<u>Preconditional Values</u>	<ul style="list-style-type: none"> - For each of these values, could you please indicate which other values are related to fostering this value through explanations? These can be values other than the ones identified in this thesis.
<u>Components of algorithmization</u>	<ul style="list-style-type: none"> - Based on the intended audience and the related values, which components of algorithmization need to be included in an explanation in order to foster each of these values?

C Interview Preparation

Dear Interviewee,

Thank you very much for agreeing to have this interview with me. In this document you will find some background information to my research which will help you understand a little bit better what my research is about and what we will discuss in our interview.

Holistic view on Predictive Policing and its Impact on Explainability

The first important thing to address is the specific view of predictive policing that is taken in my thesis. This view is holistic, meaning that it not only focusses on the predictive algorithm that is used in these practices but on the context of predictive policing practice in which this technology is used. This includes the whole process from data collection until its use by the algorithm, and the actions that are taken by the police based on the algorithms predictions. This view is chosen rather than a technology centred view because the literature review determined that a predictive policing system itself is not enough to produce a reduction in crime and that this system has to be viewed as part of a larger crime fighting strategy.

Explainability, is concerned with communicating information in a way that can be understood by the recipients of the information. In the literature review it was determined that this principle, has traditionally been focused on making sure that the underlying logic behind the output of an algorithm can be explained. As such, explainability has been technology centred. However, because I am using a holistic view on predictive policing in my thesis, I also use a broader understanding of explainability. To illustrate this, I introduce the concept of Broad Explainability, implying that explanations should not only cover information regarding the algorithm and how it produces predictions but also how these predictions are used in the larger context within which the process of predictive policing takes place. Finally, it is also important to note that Broad Explainability in this thesis will mainly be discussed from a practical point of view, meaning that this thesis is specifically concerned with Broad Explainability as *the act of providing explanations that cover the whole process of predictive policing*.

List of Important Values At Stake and their Potential Relationship with Broad Explainability

Through the literature review the following values were identified as being ‘at stake’, meaning values that are argued to be achieved through predictive policing and values that are argued to be endangered by predictive policing:

Value	Definition
<u>Explainability</u>	The extent to which information can be communicated in a way that can be understood by recipients.
<u>Efficiency</u>	The extent to which police resources are optimally allocated in spatial and temporal sense.
<u>Effectiveness</u>	The extent to which police operations are successfully reducing crime rates.

<u>Accuracy</u>	The extent to which a predictive algorithm makes correct predictions
<u>Security</u>	The extent to which predictive policing practice keeps people are safe from risks arising from potential crime.
<u>Accountability</u>	The extent to which the use of predictive policing algorithms can be assessed and consequences imposed based on those assessments.
<u>Responsibility</u>	The extent to which the duty of care for the proper use of the predictive policing algorithm has been clearly allocated.
<u>Transparency</u>	The extent to which information with regards to the whole predictive policing practice is made available.
<u>Comprehensibility</u>	The extent to which information about predictive policing practice can be understood by the recipient.
<u>Trust</u>	The extent to which people believe the predictive policing algorithm is treating them without prejudice and is working for their benefit.
<u>Fairness</u>	The extent to which a predictive policing algorithm considers everyone on the same basis.
<u>Privacy</u>	The extent to which private data used for predictive policing is secure and untraceable.

It is important to explain here why explainability was identified as a separate value at stake, even though it is also the phenomenon studied. This is because this thesis is looking at explainability from a practical point of view, meaning the act of giving explanations. The act of giving explanations, however, could potentially have an impact on explainability as a value, when understood as described above. Furthermore, concerns have been expressed that the comprehensibility of policing practice diminishes due to the introduction of predictive algorithms and many guidelines on the use of algorithms include explainability as an independent value. As such, explainability was identified as a separate value at stake in the context of predictive policing practice.

Based on the discussion of these values in the context of predictive policing practice, which resulted in the above described definitions, two potential ways in which these values could relate to Broad Explainability as a practice were identified: (1) Balancing, meaning that giving explanations should not diminish these values, and (2) Fostering, meaning that giving explanations could promote these values. In my interview with you I would like to discuss the definitions of these values and whether they have a fostering or balancing relation to Broad Explainability as a practice. In other words, whether the act of providing explanations could foster these values in some sense, or whether providing explanations could diminish or endanger these values.

Broad Explainability as a Practice

As explained, Broad Explainability as a practice is concerned with providing explanations that cover not only on how the system functions but also on how the system fits within the larger context in which it is used. The second part of the interview will be concerned with what type of explanations could foster the values described above. In order to discuss this we need to know what type of explanations could be given, and to determine this we

need a holistic lens through which to view predictive policing practice. In my thesis this lens is ‘algorithmization.’ Algorithmization refers to the process of organizational change around the introduction of algorithms and looking at predictive policing practice through this view provides a clear way of dividing the context in which predictive policing algorithms are used into different components. As such, each of the 7 components of algorithmization is a aspect of predictive policing practice that could be explained and explanations could also cover multiple components. The components are described as follows:

Component of Algorithmization	Description of the Component
Technology	The algorithm itself either as a standalone system or a system integrated into the organisational infrastructure.
Expertise	The level of expertise available in an organisation with regards to the use of algorithms.
Information Relations	The use by the algorithm of old information & information from outside sources, and the production of new information resulting in changing information relations.
Organizational Structure	(New) Departmental collaboration or organizational control structures resulting from the use of the algorithm.
Organizational policy	Policies surrounding the algorithm pertaining to e.g. transparency, responsibility and maintenance.
Monitoring and evaluation	Methods of monitoring and evaluating foreseen and unforeseen consequences of the use of the algorithm.
Socio-Technical Relations	The interplay between the outputs of an algorithm and human decision making with regards to these outputs which result in certain actions being taken.

In this second part of the interview will build on the results of the first part of the interview as we will discuss the values which you indicated are most important to foster in the context of predictive policing practice. Per value, we will discuss: (1) who the intended audience of an explanation fostering this value could/should be, (2) which other values are related to this value, and finally, based on this knowledge (3) which components of algorithmization should be included in the explanation.

Thank you once more for agreeing to participate in the upcoming interview. I hope to have informed you enough about the content of the interview and I look forward to having a fruitful meeting with you.

Kind regards,
Tim Vrieling

D Fill In Document

Value	Definition	Relationship (Balancing/Fostering)
<u>Explainability</u>	The extent to which information can be communicated in a way that can be understood by recipients.	
<u>Efficiency</u>	The extent to which police resources are optimally allocated in spatial and temporal sense.	
<u>Effectiveness</u>	The extent to which police operations are successfully reducing crime rates.	
<u>Accuracy</u>	The extent to which a predictive algorithm makes correct predictions	
<u>Security</u>	The extent to which predictive policing practice keeps people are safe from risks arising from potential crime.	
<u>Accountability</u>	The extent to which the use of predictive policing algorithms can be assessed and consequences imposed based on those assessments.	
<u>Responsibility</u>	The extent to which the duty of care for the proper use of the predictive policing algorithm has been clearly allocated.	
<u>Transparency</u>	The extent to which information with regards to the whole predictive policing practice is made available.	
<u>Comprehensibility</u>	The extent to which information about predictive policing practice can be understood by the recipient.	
<u>Trust</u>	The extent to which people believe the predictive policing algorithm is treating them without prejudice and is working for their benefit.	
<u>Fairness</u>	The extent to which a predictive policing algorithm considers everyone on the same basis.	
<u>Privacy</u>	The extent to which private data used for predictive policing is secure and untraceable.	

Value	Audience	Related Values	Components of Algorithmization

Component of Algorithmization	Description of the Component
Technology	The algorithm itself either as a standalone system or a system integrated into the organisational infrastructure.
Expertise	The level of expertise available in an organisation with regards to the use of algorithms.
Information Relations	The use by the algorithm of old information & information from outside sources, and the production of new information resulting in changing information relations.
Organizational Structure	(New) Departmental collaboration or organizational control structures resulting from the use of the algorithm.
Organizational policy	Policies surrounding the algorithm pertaining to e.g. transparency, responsibility and maintenance.
Monitoring and evaluation	Methods of monitoring and evaluating foreseen and unforeseen consequences of the use of the algorithm.
Socio-Technical Relations	The interplay between the outputs of an algorithm and human decision making with regards to these outputs which result in certain actions being taken.

E Overview Interviews






























Expert	Date	Duration
Preliminary Interviews		
<u>Expert 5</u>	15-06-2021	52:54
<u>Expert 0</u>	21-06-2021	54:52
Main Sample Interviews		
<u>Expert 1</u>	29-06-2021	01:03:38
<u>Expert 2</u>	30-06-2021	01:02:34
<u>Expert 3</u>	02-07-2021	39:40
<u>Expert 4</u>	05-07-2021	57:54
<u>Expert 5</u>	05-07-2021	50:58
<u>Expert 6</u>	05-07-2021	56:57
<u>Expert 7</u>	06-07-2021	01:01:50
<u>Expert 8</u>	06-07-2021	33:27
<u>Expert 9</u>	09-07-2021	01:01:07
<u>Expert 10</u>	21-07-2021	51:55

F Sample Description

Expert	Academic Background	Type of Knowledge (Academic, Practical, Both)	Relationship to predictive policing	Sample Selection Method
0	Professor of Governance and Innovation	Academic	Worked since 2014 on the subject of Big Data. In this context also researched into predictive policing in the European context. Focus area is ethics surrounding the use of data in governance.	Google Search
1	LM in Law, currently working at the Institute of Criminology in Ljubljana, Slovenia	Both	Has been conducting research on the intersection of criminal law and technology for 8-9 years. Did a specific study in predictive policing, in the context of the Slovenia's legal system.	Own Extended Network (Convenience)
2	PhD in Sociology and a MA in Technology Studied	Academic	Currently part of the CUPP project which does research into the topic of predictive policing.	Own Extended Network (Convenience)
3	Director of the Laws, Technology & Society Research Group.	Academic	Has been working on AI and the use of AI by law enforcement. Mainly engages with the topic of predictive policing from the perspective of fundamental rights – especially data protection as a fundamental right.	Own Extended Network (Convenience)
4	PhD Candidate doing research on the Dutch CAS system – a place-based predictive policing system.	Both	Has been working on her PhD for three years., during which she also spent 2 years in the field with the Dutch police.	Own Extended Network (Convenience)
5	Assistant Professor in Criminal Law and Philosophy. Background in applied ethics.	Academic	Is currently mainly doing research on the intersection between ethics, law and technology – of which predictive policing is a prime example.	Own Extended Network (Convenience)
6	Criminology background. PhD on the police's use of ICT.	Both	Has been doing research on the area of policing and ICT for nearly 20 years. Previously affiliated with the Police University College in Norway.	Own Extended Network (Convenience)
7	Senior Researcher at the Centre for Security Studies in Zurich. Background in political science and	Both	Research agenda over the past 10 years has emerged along the lines of technology and policing. Has done in depth research on predictive policing in Switzerland.	Google Search

	international relations.			
8	PhD is a combination between psychology and AI.	Both	Has worked for TNO (Dutch Research Centre for Applied Science) where she worked for the police. Currently works at Xomnia, a company that has developed several predictive policing tools as an analytics translator.	Google Search
9	Currently works for TNO.	Both	Does research for the police. Has conducted research on the Dutch CAS system in the past.	Snowball Sampling
10	Journalist at Algorithm Watch	Academic	Has been writing on the topic of AI for two years. Has previously written articles on predictive policing.	Own Extended Network (Convenience)

G Coding Scheme

●  Code System	1416
▼ ●  General Comments	0
●  Predictive Policing Context	8
●  Broad Explainability as a Practice	36
●  Values	22
●  Definitions	17
▼ ●  Interview Part 1	0
▼ ●  Values from the Benefits of Predictive Policing	0
> ●  Efficiency	90
> ●  Effectiveness	62
> ●  Accuracy	54
> ●  Security	58
▼ ●  Values from the Concerns with Predictive Policing	0
> ●  Explainability	118
> ●  Accountability	122
> ●  Responsibility	94
> ●  Transparency	134
> ●  Comprehensibility	55
> ●  Trust	121
> ●  Fairness	117
> ●  Privacy	86
▼ ●  Interview Part 2	0
▼ ●  Most Important Values to Foster	2
●  Important Values	11
●  Reason for Importance	7
▼ ●  How to Foster Them	0
●  Audience	54
●  Related Values	90
●  Components of Algorithmization	58

Declaration of Authorship

I hereby declare that, to the best of my knowledge and belief, this Master Thesis titled **“Using Explanations to Foster Values: Expert Opinions on the Potential of Broad Explainability to Foster Different Values in the Context of Predictive Policing”** is my own work. I confirm that each significant contribution to and quotation in this thesis that originates from the work or works of others is indicated by proper use of citation and references.

Tallinn, 09 August 2021

Tim Wei Shi Vrieling

Consent Form

for the use of plagiarism detection software to check my thesis

Name: Vrieling

Given Name: Tim Wei Shi

Student number: KU Leuven: R0781074 | WWU Münster: 509675 | TalTech:
195264MVGM

Course of Study: Public Sector Innovation and eGovernance

Address: Stoepveldsingel 95, 9403SM, Assen, Netherlands

Title of the thesis: Using Explanations to Foster Values: Expert Opinions on the Potential of Broad Explainability to Foster Different Values in the Context of Predictive Policing.

What is plagiarism? Plagiarism is defined as submitting someone else's work or ideas as your own without a complete indication of the source. It is hereby irrelevant whether the work of others is copied word by word without acknowledgment of the source, text structures (e.g. line of argumentation or outline) are borrowed or texts are translated from a foreign language.

Use of plagiarism detection software. The examination office uses plagiarism software to check each submitted bachelor and master thesis for plagiarism. For that purpose the thesis is electronically forwarded to a software service provider where the software checks for potential matches between the submitted work and work from other sources. For future comparisons with other theses, your thesis will be permanently stored in a database. Only the School of Business and Economics of the University of Münster is allowed to access your stored thesis. The student agrees that his or her thesis may be stored and reproduced only for the purpose of plagiarism assessment. The first examiner of the thesis will be advised on the outcome of the plagiarism assessment.

Sanctions. Each case of plagiarism constitutes an attempt to deceive in terms of the examination regulations and will lead to the thesis being graded as "failed". This will be communicated to the examination office where your case will be documented. In the event of a serious case of deception the examinee can be generally excluded from any further examination. This can lead to the exmatriculation of the student. Even after completion of the examination procedure and graduation from university, plagiarism can result in a withdrawal of the awarded academic degree.

I confirm that I have read and understood the information in this document. I agree to the outlined procedure for plagiarism assessment and potential sanctioning.

Tallinn, August 9, 2021

Tim Wei Shi Vrieling