Discrimination-Aware Data Analysis for Criminal Intelligence

Pragya Paudyal

A thesis submitted to Middlesex University in partial fulfilment of the requirements for the degree of Doctor of Philosophy

School of Engineering and Information Sciences

Middlesex University

December 2019

Candidate's declaration form

Middlesex University
Research Degree

Middlesex University London

Candidate Declaration form

Student details

Student name:	Pragya Paudyal
Student number:	M00335089
Thesis Title:	Discrimination-Aware Data Analysis for Criminal Intelligence Analysis
Degree for which thesis is submitted	PhD

Candidate Declaration

1 Research Integrity

I declare that the work presented is wholly my own, unless clarified as part of the submission.

2 Material submitted for another award

either	*I declare that no material contained in the thesis has been used in any other submission for an academic award
or	~~*I declare that the following material contained in the thesis formed part of a submission for the award of .....(state awarding body and list the material below)~~

*delete as appropriate

3 Research Ethics

I confirm that the research submitted has been subject to ethical review and has not deviated from the terms of ethical approval given by the Research Ethics Committee.

Ethics ID number:..... (this can be found in the MORE system or in your ethics approval letter)

Statement by the Student

Signature of Student: Pragya	Date: 27th November 2019

1 Registration from V4_28-11-16

Abstract

The growing use of Machine Learning (ML) algorithms in many application domains such as healthcare, business, education and criminal justice has evolved great promises as well challenges. ML pledges in proficiently analysing a large amount of data quickly and effectively by identifying patterns and providing insight into the data, which otherwise would have been impossible for a human to execute in this scale.

However, the use of ML algorithms, in sensitive domains such as the Criminal Intelligence Analysis (CIA) system, demands extremely careful deployment. Data has an important impact in ML process. To understand the ethical and privacy issues related to data and ML, the VALCRI (Visual Analytics for sense-making in the CRiminal Intelligence analysis) system was used . VALCRI is a CIA system that integrated machine-learning techniques to improve the effectiveness of crime data analysis. At the most basic level, from our research, it was found that lack of harmonised interpretation of different privacy principles, trade-offs between competing ethical principles, and algorithmic opacity as concerning ethical and privacy issues among others.

This research aims to alleviate these issues by investigating awareness of ethical and privacy issues related to data and ML.

Document analysis and interviews were conducted to examine the way different privacy principles were understood in selected EU countries. The study takes a qualitative and quantitative research approach and is guided by various methods of analysis including interviews, observation, case study, experiment and legal document analysis.

The findings of this research indicate that a lack of ethical awareness on data has an impact on ML outcome. Also, due to the opaque nature of the ML system, it is difficult to scrutinize and as a consequence, it leads to a lack of clarity in terms of how certain decisions were made. This thesis provides some novel solutions that can be used to tackle these issues.

This thesis is dedicated to my father and mother
who always worked hard to give us better education and better future.

Acknowledgements

Undertaking this PhD was possible with the support of many people including my supervisors, family and friends, and I would like to express my sincere gratitude to all of them. First and foremost, I would like to thank my supervisors: Prof. William Wong and Dr. Leishi Zhang for their guidance, extensive academic support and consistent encouragement throughout my research work. Thanks must also be given to Dr Neesha Kodagoda for giving guidance and encouragement. William, Leishi and Neesha, I am really indebted for your important feedback, words of inspiration and cooperation throughout my research.

I would like to thank Dr Penny Dequony, who was my second supervisor and retired during my research work. I really appreciate her contributions, her precious time and ideas during my initial research. I further like to thank my peers and members of the School of Science and Technology I met during my PhD who have extended their support to make this thesis possible, Dr Chris Rooney, Dr Simon Attfield, Pirkko Harvey and Nallini Selvaraj. I would like to thank my friends, Suzanne, Mithila, Gayathri, Karthika, Celeste and Poujha for their support and being there when I need them.

I would further like to extend my deepest appreciation to Ms Joanna Loveday for constructive criticism of the thesis.

Nobody has been more important to me in the pursuit of this project than the member of my family. I would like to thank my father-in-law Paras, mother-in-law Debi, uncle Dr Jagannath and aunt Laxmi. They have encouraged and helped me at every stage of my personal and academic life and have wished to see this achievement come true. I would like to thank my family in Belgium, Nepal and the UK for their support, well wishes, words of inspiration and appreciation throughout my research.

Most importantly, I wish to thank my father Chiranjibi, mother Bishnu, brother Bharat, sister-in-law Binita, niece Adhbhika and husband Bishesh for their unfailing love, support and continuous encouragement, not only during the years that it has taken to write this thesis, but also for their constant cooperation and motivation of my academic pursuits over the years.

Acknowledgement of funding

The research described in this thesis was funded by the Computer Science Department of Middlesex University, and was linked to the research carried out from the European Union 7th Framework Programme FP7/2007-2013, through the VALCRI project (2014-2018) under grant agreement no. FP7-IP-608142, awarded to B.L. William Wong, Middlesex University London, and Partners.

Publications related to this thesis

Paudyal, P., & William Wong, B. L. (2018). Algorithmic Opacity: Making Algorithmic Processes Transparent through Abstraction Hierarchy. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 192–196. https://doi.org/10.1177/1541931218621046

Paudyal, P., Rooney, C., Kodagoda, N., Wong, B. L. W., Duquenoy, P., & Qazi, N. (2017). How the Use of Ethically Sensitive Information Helps to Identify Co-Offenders via a Proposed Privacy Scale: A Pilot Study. In *2017 European Intelligence and Security Informatics Conference (EISIC)* (pp. 164–164). IEEE. https://doi.org/10.1109/EISIC.2017.35

Marquenie, T., Coudert, F., Duquenoy, P., & Paudyal, P. (2017). Roadmap for the resolution of ethical and human rights issues in automated data analysis and extraction computations in VALCRI. VALCRI White Paper Series. Retrieved from https://lirias.kuleuven.be/1711906?limo=0

Table of contents

List of figures

List of tables

List of abbreviations

ACM- Association for Computing Machinery

AH- Abstraction Hierarchy

AI- Artificial Intelligence

CIA- Criminal Intelligence Analysis

COMPAS- Correctional Offender Management Profiling for Alternative Sanctions

EU- European Union

GDPR- General Data Protection Regulation

GPS- Global Positioning System

ICO- Information Commissioner's Office

IEB- Independent Ethics Board

KDD - Knowledge Discovery from Data

kNN - k nearest neighbour

ML- Machine Learning

MDS – Multi-Dimensional Scaling

PCA - Principal Component Analysis

PDP- Partial Dependence Plots

PII- Personally Identified Information

VALCRI- Visual Analytics for sense-making in CRiminal Intelligence Analysis

WDA- Work Domain Analysis

Chapter 1

Introduction to the study

1.1 Synopsis

We live in a society where algorithms are increasingly being used to assist certain tasks, and sometimes replace human intervention in the decision-making process. The use of algorithms has helped in analysing vast amounts of data, that would otherwise have been impossible for a human to do manually. However, algorithms operate like black boxes, which work without explaining the outcome of certain decisions. The rising ubiquity of algorithmic use in decision-making in society raises several fundamental questions concerning its transparency, privacy and ethical issues. The goal of this thesis is to (i) understand some of the privacy and ethical concerns that could impact on developing a Criminal Intelligence Analysis (CIA) system, and (ii) provide some new solutions that can be used to tackle the issues using various approaches which includes analysing different privacy and legal systems, data sensitivity measurement and use of different visualisation techniques.

1.2 Research background

In the era of big data, an increasing volume of data is being collected from a range of different sources, and at a greater speed than ever before. The complexity and amount of data available today challenges traditional analytical approaches in data science and data analytics. In this context, algorithms have played an important role in making sense of large quantities of potentially dynamic data. An algorithmic approach can process a large volume of data to find out the patterns that subsequently become useful in the decision-making process (Royal Society n.d.). As a result, these approaches have been used in different domains such as hiring (Chalfin et al. 2016), policing (Couchman 2019), criminal intelligence (Jester, Casselman and Goldstein 2015), and health (Starr 2018) among others.

The growing use of algorithmic approach presents opportunities, as well as ethical and social issues. Cathy O'Neil in her book "Weapons of Math Destructions" states *"...mathematical models can sift through data to locate people who are likely to face great challenges, whether from crime, poverty, or education. It's up to society whether to use that intelligence to reject and punish them—or to reach out to them with the resources they need*". Scholars and activists have pointed to a range of social, ethical and legal issues associated with machine learning decision-making including bias and discrimination (Williams, Brooks and Shmargad 2018), lack of transparency (Barocas and Selbst 2016; Pasquale 2015; Stefaan Verhulst, 2015), fairness (Josephson, 2016; O'Neil, 2016) and accountability (Diakopoulos, 2014; Mittelstadt et al. 2016; Stahl, Timmermans and Mittelstadt 2016).

The use of algorithms in different domains is facilitated by (i) big data, and (ii) advancement of increasingly sophisticated machine learning algorithms.

1.3 Big data

We are living in an age where we generate data through various things that surround us in our everyday lives. We have many devices such as a smartwatch, mobile, tablets among others which are connected and are collecting data. Big data is often referred to as the "three Vs". These are the increasing Volume of data, the increased Velocity with which it is produced and processed, and the increased Variety of data types and sources. These data can be used for analysis purposes to monitor disease outbreaks, evaluate credit risks, improve retail and manufacturing, pattern recognition in medical diagnosis , understand the increasing crime type and a particular characteristic that is associated with high reoffending rates among others. According to Moses and Chan (2014) the size and complexity of datasets being analysed are important to gain insight into the data. However, big data analysis is not possible using the traditional statistical, or any empirical techniques. But the recent advancement of sophisticated machine learning algorithms has made it possible to analyse big data.

1.4 Algorithmic system

To make sense of these big data, we need an algorithmic system. Machine learning algorithms is a subdomain of Artificial Intelligence (AI), commonly used to analyse the big data. According to Van Otterlo (2013) machine learning is "*any methodology and set of techniques that can employ data to come up with novel patterns and knowledge, and generate models that can be used for effective predictions about the data*". Machine learning allows a computer to learn directly from example, data and experience (Royal Society n.d.). In some specific domains, machine learning systems are already able to achieve a higher level of performance than humans. For example, in one image labelling challenge, the accuracy of machine learning has increased from 72% in 2010 to 95% in 2015, exceeding human accuracy (The Economist 2016). Traditional approaches rely on hard-coded rules, which set out how to solve a problem, step-by-step. In contrast to this, machine learning systems are given a large amount of data to use as an example of how the task can be achieved or from which to detect patterns, and tasks are set. Based on the example, the system then learns how best to achieve the desired output. The algorithm defines decision-making rules to handle new inputs. Significantly, the human operator does not need to understand the rationale behind the decision-making rules defined by the algorithms (Matthias 2004). This grants algorithm some degree of autonomy, consequently, tasks performed by the machine learning system are difficult to predict beforehand or explain afterward. As a result, transparency is required to understand the logic and reasoning behind certain outcomes. However, the rationale of the algorithms working is obscured, lending to the portrayal of machine learning algorithms as "black box". The problem of algorithmic opacity or the "black box" nature of the algorithms presents different ethical and legal issues.

1.5 Ethical and social implication of using algorithmic system for decision-making process

Algorithmic decisions can mimic and amplify patterns of discrimination, due to the models being trained on historical data. This includes decision-makers' prejudice or reflect the bias present in the society (Barocas and Selbst 2016). O'Neil (2016) in her book "Weapons of

Math Destruction" highlights several case studies, particularly in the areas of criminal intelligence, on harm and risk of public accountability associated with data-driven algorithmic decision-making. A recent study by ProPublica of proprietary algorithms COMPAS, used by courts in the US to calculate recidivism (Larson et al. 2016), the probability of someone committing a crime again once they are out of jail based on their past crime data found that the algorithm was biased towards black defendants compared to white defendants. Also, the outcome produced by the system was invariably inaccurate. Lack of transparency and little oversight of the inner workings of a system can erode the rule of law and diminish individual's rights.

Building an ethical machine learning algorithm is difficult and challenging as machine learning is the technology that allows the system to learn directly from examples and data rather than following pre-programmed rules. Machine learning algorithms do not reason as humans do, and this makes their outcome difficult to predict and explain. When machine learning algorithms are being used in a critical domain such as policing or medicine, it is important to consider the ethical, legal and social implications and minimize or eliminate it. Currently, revolutionary technology has outpaced ethical policies, as a result, our policies have not kept pace with technological development.

Fortunately, there is increasing awareness of algorithmic detrimental effects on different ethical, legal and social effects and of the need to reduce or eliminate them. In addition to this, journalists (Diakopoulos 2014), politicians (House of Common 2018; The White House 2016), and researchers have expressed the need for technique/s to assess whether decisions produced by algorithms conform to ethical standards.

1.6 Research motivation

As recent literature, research and news demonstrate, algorithmic decision-making tools are being used in sensitive and complex domains such as policing and criminal intelligence. We are living in an era where our data are being collected, documented, processed and interpreted without us knowing how our data are being used. There are many different technologies and these technologies are used for collecting and generating data. These

technologies are not just producing large volumes of data but also different types of data such as structured, unstructured, video, etc. It would be difficult for a human to sift through all the necessary data to make a decision. As a result, the number of domains and people using machine learning for decision-making will increase, and so will the development of these technologies. The lack of awareness of the ethical, legal and social issues can have a huge impact on civil liberties and human rights.

When any new technology is used, especially in the area of criminal justice, we aim to improve the security of citizens without infringing the freedom of individuals or society. However, as the complexity of these technologies is often hidden from users, it is difficult to understand and verify the correctness of these technologies. The ethical principles of 'do no harm' and 'justice' are at stake. Analysts cannot verify the accuracy of analysis due to the complexity of the algorithmic process. As a result, sentencing can be based on socio-cultural demographics of an individual instead of the nature and severity of the crime (Gov.uk 2019; Larson et al. 2016; O'Neil 2016). This is an emerging and new research area and the issues will increase and combine with the novelty of machine learning to contribute to a general picture of confusion as far as machine learning is concerned. At present, there is an absence of ethics in the design of these artefacts. Ethics is needed where decisions made by machine affects individuals and society as a whole.

The research carried out in this thesis was part of the EU-funded project VALCRI (Visual Analytics for sense-making in Criminal Intelligence analysis) FP7-IP-608142. It involved multiple partnerships between researchers working in different disciplines (e.g. Software engineering, Machine learning, HCI, Ethical, Legal, Security and Privacy). End-user partners for the VALCRI project were from 3 police forces from Belgium and UK, and a data protection partner from Germany. VALCRI was designed and developed for law enforcement agencies as a next-generation criminal intelligence analysis and investigation system based on a sense-making technology supported by advanced data processing and analytics software. VALCRI integrated machine-learning techniques and interactive visualisation methods to improve the effectiveness of crime data analysis.

Machine learning algorithms may appear to be morally neutral but there are many growing concerns about machine learning and ethics which has an impact on individual's lives. From the beginning of the project, ethics specialists were embedded in the project and externally

through the establishment of an Independent Ethics Board (IEB). The IEB and project researchers identified several ethical concerns (IEB 2017): accidental discrimination, the mosaic effect, algorithmic opacity, data aggregation with mixed levels of reliability, data and reasoning provenance, and various biases.

VALCRI consists of several work packages. Issues related to bias, data aggregation, and provenance were addressed by other teams. This thesis focuses on (i) **accidental discrimination** occurs when the outcome of neutral system processes results in the unexpected discrimination and unfair treatment of certain groups or people, for instance, recent news about gender discrimination against the Apple card. The algorithm used to set credit limits for the new Apple card were biased towards male. Apple's credit offered different credit limits for husband and wife despite their having no separate bank accounts or separate assets. (ii) **mosaic effect** occurs when harmless isolated data are combined and used to paint a clearer and encompassing picture of who an individual is and what he or she likes such as the combining the netflix dataset and IMDB datasets by which researcher were able to uncover apparent political preferences and other potentially sensitive information **and (iii) Algorithmic opacity** takes place due to the complexity and 'black box' nature of the algorithmic process for example, not being aware of the input and correlation of the input due to the opacity of the algorithmic process.

1.7 Research questions and objective

This thesis addresses the following research question:

How can the discrimination- aware approach support analysts making ethical decisions in criminal intelligence?

Discrimination during the analysis process is often considered to originate from using specific attributes such as gender, ethnicity etc. As a result, prior to the analysis process, this information is removed to produce fair, indiscriminate and unbiased results. However, this does not always solve the problem. From some of the earlier research it was found that, even removal of this information leads to a discriminatory or biased outcome and low accuracy

rate, although analysts thought the outcome was fair, accurate and did not have any ethical issues

However, our concept of discrimination-aware is to use this information during the analysis process so that the analyst is aware of these features are being used. A discrimination-aware approach will help the analyst to be aware of any ethical or privacy issue during the analysis process. To produce an effective discrimination aware approach, a researcher must have some conception of what the aspects to focus on are.

Therefore, the aims of this research are:

1) To identify the requirements for compliant data processing according to the EU data protection law and how it is interpreted in different EU countries (Belgium, Germany and the UK) (Chapter 3). Data is an important aspect in machine learning, when collecting and using data (especially ethically sensitive data), we need to respect legal and privacy considerations.
2) To identify the data within the police domain that are considered as ethically sensitive data. The Privacy scale was proposed to compare between different EU countries, the ethical sensitivity of data within the policing domain (Chapter 4).
3) To understand the complexity of machine learning models using the Abstraction Hierarchy (AH), AH is a method to describe a system on several level of abstraction. The AH is a commonly used framework in cognitive engineering to visually represent the relationships between low level objects in a process and their relationships with various higher order functions and goals of the process. In our research, the AH approach was used to study the effect of ethically sensitive information on the higher order goals of (i) Information Flow and (ii) Information Accuracy (Chapter 5).
4) To understand the effect of removing ethically sensitive information from a criminal's record during the data analysis process to assess the criminal's likelihood to re-offend and aid analyst in understanding the analysis process using some visualization techniques (Chapter 6).

1.8 Research methodology

This research adopts a mixed-methods approach to the study. Mixed methods are a procedure for collecting and analysing both qualitative and quantitative data within the same study to gain a better understanding of the research problem (Tashakkori and Teddlie 2003). An important aspect of any research is the traceability of the steps taken in arriving at the final products. Therefore, the choice of methodology is an important part of the research process, as it provides a suitable framework and influences the study design, data collection, analysis methods and findings presented (Creswell 2013).

Study 1: Comparative analysis of "The interpretation of the principle of purpose limitation" across the EU: Belgium, Germany and the UK

The aim of this study is to understand how the term "purpose limitation", a key data protection principle which requires the collection and processing of data to have a clear defined purpose, - including the key concepts within the term is interpreted and operationalized in different EU countries (Chapter 4). As the police end-users for VALCRI were from Belgium and the UK and our data protection research partner was from Germany therefore these countries were used for the comparative analysis.

A qualitative approach was used for this study. Data was collected over three phases for this study. The first phase study was conducted using secondary data sources. Secondary data sources involved reviewing several legal documents from the EU directive: Belgium, Germany and the UK. After the data collection, common themes based on the purpose limitation principle were identified using open coding method. Four common themes (specified, explicit, legitimate purposes, and compatible use) from the EU directive: Belgium, Germany and the UK were identified. Based on the four themes, commonalities and divergences between selected EU countries were noted. Some of the information was not available in the legal documents, so second phase study was conducted. The second phase study involved observations collected while listening and taking notes during several VALCRI consortium meetings, work package meetings, and meetings between police end-users (UK and Belgium) and data protection (Germany) and legal (Belgium) experts from different countries. Once the comparative analysis was created, to evaluate the content

involved collaborative work between researchers on the VALCRI project, who are in the legal field from the UK and Belgium.

Study 2: Development of a prototype privacy scale

In this study, a privacy scale was developed. The study sought to understand different types of information that was considered private and therefore ethically sensitive information is in the policing domain. Using privacy scale, the study then attempted to compare how this information is considered between different countries.

The goal of this study was to gain an in-depth understanding of the way analyst understood the different privacy-relevant information (ethically sensitive information), and where they placed them on the privacy scale. Therefore, a qualitative methodology was chosen for this study. Ethically sensitive information used in the privacy scale was based on the anonymized data given by a police officer for the project. The validity of selection of data was confirmed by police analysts. Seven operational intelligence analysts participated from different police forces in the UK and Belgium. They were interviewed and were asked to give a score to rate how private they considered each piece of information they used about individuals.

Study 3: Modelling of privacy information using the abstraction hierarchy

This study aimed to visually represent the effect of privacy information on (i) information flow and (ii) information accuracy.

The complexity of the algorithmic process was visualised using Abstraction Hierarchy approach, abstraction hierarchy is a method to describe a system on several level of abstraction. A case study was used to develop a conceptual understanding of how privacy information has an impact on information accuracy for the VALCRI system.

Study 4: Discrimination aware machine learning

This study aims to understand the effect of ethically sensitive information during an assessment of the criminal's likelihood to re-offend. And also, to understand the impact of algorithm selection during the analysis process.

The study was conducted using two sets of datasets, one including ethically sensitive information and other excluding ethically sensitive information to understand the impact on accuracy. Five different classification algorithms were used. The same set of classification algorithms were used for both datasets. A quantitative methodology was used for the evaluation.

1.9 Thesis outcome (contributions)

In this thesis, I offered a narrative literature review which covers various ethical issues that occur in criminal intelligence analysis as a result of using the algorithmic process. I explored privacy principles which are relevant during the data analysis process. From the literature review, I have identified certain areas which need further research, especially related to ethically sensitive data. For that I have developed:

- A guideline which illustrates one of the important principles regarding data collection and further use within the General Data Protection Regulation (GDPR). This provides a guideline on further using data for some other purpose, as every country has a different interpretation on the further use of the data.
- A privacy scale framework, which allows the analyst to understand the different privacy-relevant information on privacy scale based on their sensitivity level.
- Visualization of the algorithmic process is to visually represent the relationships between low level objects in a process and various higher order functions and goals of the process. The aim is to provide more transparency to the machine leaning process and raise awareness on the uncertainty and bias that could be introduced in the process.

1.10 Thesis structure

This dissertation comprises seven chapters. I briefly outline what they are and explain how each chapter relates to the next as follows:

Chapter 1 presents the introduction to the research area of the study, research motivation, research question and objectives, research methodology, thesis outcome and thesis structure.

Chapter 2 provides a review of the literature pertinent to the study. The chapter includes an introduction, the core work in machine ethics, data ethics, privacy, ethical challenges and machine learning especially in the area of policing and criminal intelligence. This will be followed by discussing the current approaches to solve the ethical issue.

Chapter 3 provides a comparative analysis of privacy law to understand data ethics and privacy aspect of data collection and the use of data. In the age of big data, more and more data have been used and further analysed. Big data and machine learning techniques have changed the traditional forms of data analysis and created a new approach to analyse data into useful or interesting information. Big data often contains huge amount of personal identifiable information. Reuse of data by repurposing and using all available data often leads to privacy challenges. The principle of purpose limitation limits the process of personal data unless it serves a concrete purpose, permitted by law or with consent of the person in questions. This chapter aims to explore and compare purpose limitation in the processing of personal data. This chapter describes a study, which compares the notion of "purpose limitation" within data protection acts across certain EU countries. The countries were selected from the VALCRI project partners.

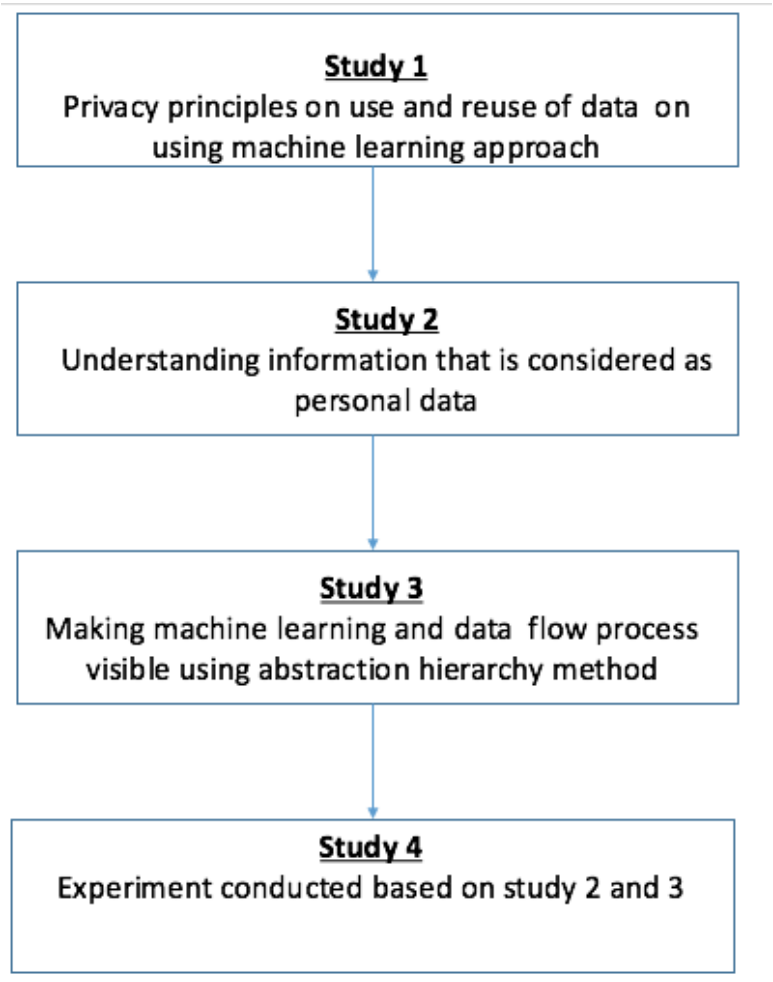
Chapter 4 presents the second study that was conducted to understand data that are considered ethically sensitive in the domain of policing. It is difficult to understand what types of data are considered as personal data, as there is no clear guideline for this. In this chapter, a list of data that are considered as personal information based on an interview and survey with a criminal analyst. Based on the list of data the information was categorised according to its impact level- privacy scale.

Chapter 5 provides a visualisation approach for understanding the ethical issue in emerging technology (using VALCRI system as a case study). Big data and machine learning complement each other one cannot exist without another. In this chapter abstraction hierarchy method was used to provide a visual form for machine learning approach. This study was

conducted to make the data and computational processes that impact ethical decision making visible.

Chapter 6 describes the final study based on the impact of using ethically sensitive information during the analysis process for more accurate results. Data analysis was conducted based on the second and third studies. Different visualisation techniques were used to help analysts to provide insight into the data use for ethical decision-making.

Chapter 7 concludes the importance and impact of a discrimination-aware data analysis approach in criminal intelligence. It puts forward recommendations for a future research study, highlights the significance of the findings and their contribution.



Chapter 2

Literature review

2.1 Introduction

The previous chapter provided an overview of the study. In this chapter, I discuss the gaps in the existing algorithmic decision-making process and its impact on individuals and society. This research is aimed at exploring and theorizing the need for ethical awareness when using algorithms in the decision-making process. Such an undertaking requires a multidisciplinary effort involving a review of literature from different domains, which forms the theoretical structure of this thesis.

To set the context of my research, the chapter begins with a review of the machine learning and its role in the decision-making process. The importance of ethics when using algorithms for the decision-making process is then discussed, including the analysis of machine ethics and data ethics, the two important factors that need to be taken into consideration in machine learning applications.

Furthermore, to address the ethical/legal issues of machine learning, I explore the underlying ethical principles as the model for ethical guidelines related to policing. The review includes the ethical issues that have been discussed in the literature. This led to the identification of approaches that have been taken to solve the problem as well as the gaps in the research.

2.2 Use of machine learning in decision-making process

We live in an era of big data, which has been integrated into every aspect of our lives. Vast amounts of data is collected, stored and processed every second. Data are collected in different forms such as unstructured, structured and formats such as text, pictures, sound, documents, video etc. Vast and disparate data are constantly and rapidly being collected and stored. As the increasingly large amount of data is collected every day, it has become a

daunting and impossible task for a human to sift through all relevant information to piece together all pieces of information to draw all possible connection. At present, data is valuable and they are the most important assets in the world (Harari 2018). These data sets provide tremendous benefits in making a variety of administrative and governmental decisions. The use of machine learning algorithms assist the extraction of meaningful information from text, documents, images and videos, detect patterns and present such data in an accessible and insightful way to its users.

An Algorithm is defined as, *"a finite sequence of well-defined instructions that describes in sufficiently great detail how to solve a problem"* (Kraemer, van Overveld and Peterson 2011). Algorithms can be understood as step by step instructions given to solve a particular problem or to make certain decisions. To make a certain decision, we will need to take some input data (selection of input), process the data (based on the instructions) and display the output.

Algorithms need data, and their effectiveness and value tend to increase as more data are used and as more datasets are brought together. Traditional rule-based systems require decision-making rules, like flowcharts, where the steps methodologies and outcome can be traced to pre-programmed instructions inputted by a human. However, due to the large volume of data, traditional rule-based approaches do not capture the desired input-output relationship well (Veale 2019). As a result, more advanced algorithmic system, such as machine learning approaches are used.

Machine learning is a "*technique that allows the computer to learn directly from examples, data and experience, finding rules or patterns that a human programmer did not explicitly specify*" (Rieke, Bogen and Robinson 2018). Machine learning algorithms are not programmed to solve a particular problem. Instead, they are programmed to learn solve problems. This technique uses a variety of algorithms that iteratively learn from data and predict outcomes. With the help of machine learning, governments are able to make better decisions and predictions by combining human and machine learning in a smart way.

Machine learning algorithms are being used in different sectors. Algorithms have brought big changes for better medical diagnoses, and assist police in decision-making (House of Common 2018). As a result, operations and decisions, which were previously performed by

a human, are increasingly delegated to algorithms (Mittelstadt et al. 2016). Algorithms mediate in the criminal justice system, credit market, higher education, employment and social media to decide for us, about us or with us in ways that has an impact in our economic and social lives.

Algorithms in the policing domain are used in a wide range of ways including predictive crime mapping, risk assessment, video and photographic analysis, and DNA profiling among others. Use of algorithm in the justice system is not new. Algorithms such as risk assessment and profiling have been used by public agencies and considered in regulation (Veale 2019).

Depending on the nature of the problem being addressed, there are different approaches which can be used within machine algorithms. Since learning involves an interaction between the learner and the environment, we can divide learning according to the nature of the interaction. Machine learning approaches can be classified into supervised, unsupervised and reinforcement learning.

2.2.1 Supervised learning

Supervised learning is also called learning from exemplars. It constructs a predictive model by learning from training examples, where each training dataset consists of a **feature vector** (of instances) describing the event and a **label** indicating the ground-truth output (Hurwitz and Kirsch 2018). The goal is to train a model that predicts the label as correctly as possible when given new instances where the labels are unknown.

Supervised learning can be classified into classification and regression. When the output has a finite set of values or discrete outcomes, it is known as classification. For example, a classifier can be used to classify a criminal as high, medium or low-risk based on given information. In contrast, when the label is continuous it is known as regression. For instance, predicting weather temperatures based on historical patterns and current conditions to provide a prediction of the weather.

2.2.2 Unsupervised learning

Unsupervised learning involves training of machines using information that is not classified or labelled. This learning allows the algorithm to act on the information without guidance. Unsupervised algorithms look for naturally occurring patterns in the data. It is useful in exploratory analysis such as discovering the underlying structure of the data and identifying outliers. Also, unsupervised learning is commonly used when we do not have labelled datasets. For example, clustering is a commonly used unsupervised learning technique for identifying crime hot spots, especially within policing (Cheng 2018).

Clustering and dimensional reduction are some of the common unsupervised machine learning techniques used in the data analysis process. Clustering involves grouping data by similarities, dimensional reduction involves compression of the data while maintaining its structure and usefulness. Clustering helps in identifying patterns such as natural grouping or outliers in the data. It is relatively difficult to measure the performance of unsupervised algorithms. Performance measures are often subjective and domain-specific when using unsupervised learning.

2.2.3 Reinforcement learning

Reinforcement learning (RL) is a behavioural learning model (Hurwitz and Kirsch 2018). RL "*is learning what to do, how to map situations to actions so as to maximize a numerical signal*" (Sutton and Barto 2017). Typically, RL is composed of two components, an agent and an environment. The agent represents the RL algorithms, which replace the human operator, while the environment refers to the object that the agent is acting on. The agent is told about the goal but is not told which actions to take. RL enables an agent to learn in an interactive environment how to behave from its own actions and experience through trial and error. The mainstream RL algorithms are used to solve games.

Machine learning algorithms provide many benefits in analysing data to find patterns and relations (Lee, 2019; Rigano, 2018; Timothy Revell 2017). In some of the cases, machine learning attains better than human-level performance (Silver et al. 2017). In contrast to this, these technologies are also affecting the lives and rights of individuals in significantly negative ways in terms of privacy, fairness, autonomy, bias, accountability, accuracy and

discrimination (Centre for Internet and Human Rights 2015; Gillespie 2012; O'Neil 2016; Wagner 2016; Ziewitz 2015). Some authors (Diakopoulos 2014; Kraemer et al. 2011; O'Neil 2016; Wagner 2016) have argued that algorithms are inescapable and they can consist of personal and strong views based on religion or culture. Machine learning algorithms raise several significant ethical and privacy problems such as discrimination, bias, lack of transparency, and the mosaic effect among others. Before discussing the ethical and privacy problem, we will discuss the role of ethics in machine learning algorithms.

2.3 Ethics in machine learning

"Ethics" have many interrelated definitions. At the most basic, ethics is defined as the study of right and wrong; what is right to do in a given or a certain situation (Stahl, Timmermans and Mittelstadt 2016). The study of ethics in the area of machine learning is relatively new and is not fully established yet. The sheer diversity and far-reaching nature of machine learning applications in our daily life requires discussion concerning the ethical aspect of these technologies.

A number of ethical issues have been discussed and debated recently in the field of machine learning and Artificial Intelligence (AI). One of the pioneers of machine ethics, James Moor, has been discussing this topic since 1985 (Moor 1985). He warned of the dangers implicit in a machine that "acted" faster than we could react and the complexity of these technologies beyond our understanding.

When we develop any system, ethics is a key component of that system, which can determine the acceptance of the technology developed as well as legislative and other responses to the developed technologies (Stahl, Timmermans and Mittelstadt 2016). With the increasing popularity of the rapid use of machine learning for the decision-making processes, developers need to be made aware of the social and ethical consequences for a better understanding of different emerging issues.

2.3.1 Machine ethics

Machine ethics, also known as machine morality, artificial morality or computational ethics, is concerned with implementing moral decision-making faculties in computers and robots (Allen, Wallach and Smit 2006). The term 'Machine ethics' existed since the birth of Artificial Intelligence (AI). However, it is only in recent years that people acknowledged the need for ethics in the design and development of the technology.

The purpose of machine ethics has been expressed in different ways. Some have expressed that the ultimate goal of machine ethics is to create machines that can follow ethical principles (Allen et al. 2006); a machine which has an "ethical dimension"; or ensuring the behaviour of the machine is ethically acceptable to human users (Anderson and Anderson 2007). Moor (1985) has made a fine-grain distinction between different kinds of ethical machines: implicit ethical agents, explicit ethical agents and full ethical agents which a machine ethics might pursue. According to Moor, implicit ethical agents are machines that are being programmed to behave ethically or at least avoid unethical behaviour. These machines are constrained in its behaviour by its designer who is following ethical principles. In contrast to this, explicit ethical agents are able to calculate the best action in ethical dilemmas using ethical principles. A machine is considered as a fully ethical agent, if it can make an explicit ethical judgment and can reasonably justify the decision. According to Anderson and Anderson (2007) the ultimate goal of machine ethics is to create a machine that is an explicit ethical agent. However, due to the recent growth in use of machine learning in decision-making, there is a need for full ethical agents especially in sensitive domains such as medicine, criminal intelligence among others. These domains demand some forms of explanation or reasoning for humans to understand the motive behind certain outcome.

2.3.2 Data ethics

Data ethics is an emerging branch of applied ethics which describes the moral problems related to the generation, processing, sharing and using of data (Floridi and Taddeo 2016). Data ethics relates to good practices around how data are collected, used and shared. Collection and storage of a large amount of personal data is in itself a risk to individual privacy (Hasselbalch and Tranberg 2017). Data protection legislation in Europe prohibits the

collection of personal data without a specific purpose and require users consent. However, living in a time of datafication our data are being collected, documented, processed, used and interpreted without our understanding. We have so many different technologies that are connected and they are either collecting or generating data. The way these data can be retrieved and used and the potential invasiveness causes problems not just in the space of ethics but also in any approach to regulating data protection and privacy law (Richterich 2018).

Data ethics is inspired by data, algorithmic systems and data-driven insight. The ability to process greater amounts, greater types of **data,** than has ever been possible due to new technologies, to do that processing we need **algorithmic** systems to make sense and insight of data, and now there are many domains, such as criminal justice, finance and medical, that are increasingly being **data-driven** by the insight that we can generate from this data.

Machine learning algorithms and AI are designed to provide insight for collected data. The effects of data practice without ethics can be diverse, discrimination, unequal opportunity, and unjust treatment. Privacy and ethics are at its core. These are the needle on the gauge of society's power balance.

2.4 Need of ethics in machine learning

As we increasingly delegate decision-making to machines, the decisions made by machines lead to many concerns due to their impacts on societies. Machine are a reliable yet malleable control mechanism. The complexity of machines is increasing day by day and this flow will continue. Moor (1985) highlights three characteristics of machines or technology that have a direct impact on ethics. These characteristics are (i) Logical Malleability (ii) Transformation and (iii) The Invisibility Factor.

Machines are logically malleable so that they can be shaped and manipulated to do any activity that can be characterised as inputs, outputs and connecting logical operations. The logic of machines can be manipulated in an infinite way through changes in hardware and software.

The concern over machine ethics is based on the fact that machines can drastically change the way we do things. Machines are deployed increasingly in everyday life from finding the shortest distance from one place to another, to which bus or tube to catch and perform well at easily understandable goals. More sophisticated systems are then used to advise on a more complex matter such as criminal justice, health care and education. People and analysts rely on these systems and increasingly delegate the decision to them.

The ethical significance of the third characteristic, invisibility factor, is that its internal operation is hidden from view. The invisibility of internal operation provides the opportunity for:

(i) invisible programming values, a machine's design and functionality reflects the values of the programmer and intended uses. This is not a malicious act of the programmer but a lack of understanding. Development is not a neutral or linear path; many possible correct choices at a given stage of development are possible. As a result, the choice of the programmer is embedded in the code. It is difficult to detect the embedded values in the programme.

(ii) the invisible complex calculation takes the form of programmes that are so complex that a programmer or user do not understand them. According to Ormand (2011) a complex system typically comprises the different instances of known and hidden interdependencies between components; output yield from a complex system is often emergent; due to this, it is difficult to know exactly which input contributes to an observed output. The complex system exhibits several defining characteristics such as feedback, strongly interdependent variables or extreme sensitivity to initial conditions. No one will have any understanding or idea of how it is giving a certain outcome.

(iii) invisible abuse is "*the intentional use of the invisible operations of a computer to engage in unethical conduct*". Invasion of an individual's right to privacy and surveillance is an example of invisible abuse. Giving personal information about an individual to third parties for advertisement without the consent of data subject is another example.

Machine learning provides a lot of benefits to organisations; enabling them to increase efficiency and make better decisions. However, it is coupled with significant ethical challenges such as fairness, privacy and respect to human rights. If we don't sufficiently look

into these challenges in the early stages, we risk hindering the innovation and benefits data science can bring to the organisation. Algorithmic systems are processing greater amounts of data. They have an increasing level of autonomy or complexity that can lead to unintended behaviours. Oversight of the algorithmic system is difficult due to many different unpredictable places and ways data is/can be used (Mittelstadt et al. 2016). When people's data are collected in different domains, such as for criminal intelligence, medical and finance, different actors can receive information about these people and use the information for whatever purpose is necessary for them.

2.5 Privacy

The uproar over the collection of data and use of different machine learning algorithm and data mining have sparked new concern about the privacy in the age of big data. The term "privacy" has varied and sometimes conflicting interpretations. Privacy in legal concepts has been referred to as the right to be left alone, data protection rights, right to control or 'own' personal data, secrecy, anonymity, and the right to respect private and family life (The Royal Society, 2015). Privacy is usually defined in terms of the individual's control over their information, and current conceptualization of privacy and data protection focused mainly on prevention of individual harm. There have been different approaches taken to protect individual privacy in the age of big data.

2.5.1 Privacy Enhancing Technologies (PETs)

Privacy enhancing technologies are technologies that embody fundamental data protection principles to help achieve compliance and data protection regulations. PETs allow users to protect their information privacy by allowing them to decide, what information they are willing to share with other people. Examples of PETs include (i) Pseudonymization, the process of processing personal data in such a way that the data can no longer be attributed to a specific data subject, (ii) anonymization, the process of removing personal identifiers data which may lead to an individual being identified, (iii) obfuscation, protecting privacy against inference attacks (The Royal Society, 2015).

2.5.2 Privacy by Design

Privacy by design is the notion of embedding privacy measures and privacy enhancing technologies (PETs) directly into the design of the technologies(Ann Cavoukian, 2013). This process involved building features at the very beginning of the processing. Organisations or companies are required to implement technical and organisational measures, at the beginning of the design of the technology, in such a way that safeguards privacy and data protection principle of individual from the start. This approach characterized by proactive rather than reactive measures. Privacy by design is integral to the General Data Protection Regulation (GDPR) for data controller and processors, making data protection obligatory by default (Information Commissioner's Office, 2017).

2.5.3 Privacy preserving machine learning

Privacy preserving machine learning, including rapid advances in cryptography and statistics, provides powerful new ways to maintain anonymity and safeguard individual privacy. Traditional methods rely on removing identifiable information, encrypting data, and limited sharing of data. Unfortunately, this alone may be inadequate for the emerging big data systems. Privacy preserving machine learning combines complementary technologies to address these privacy challenges. Approaches such as (i) differential privacy, which involves adding mathematical noise to personal data, protecting individual privacy but enabling insight into group patterns, (ii) secure multi-party computation, where many people can combine their input to compute a function, without revealing their input to each other, (iii) homomorphic encryption, which enables computation on encrypted data without leaking any information about the underlying data (Al-Rubaie & Chang, 2018). The field of privacy preserving machine learning has emerged as a flourishing research area.

2.6 The emergence of ethical principles in machine learning

To reap the societal benefits of machine learning, we have to make sure these systems follow ethical principles, moral values and social norms that we human would follow in the same scenarios. In the field of machine learning, there is widespread agreement on some of the core issues (such as discrimination) and values (such as fairness) that machine learning algorithms should focus on. Ethical principles are designed to provide guidelines for making

a responsible and acceptable decision. Principles help condense complex ethical issues into a few important factors which can be clearly understood and agreed upon by people from different fields and domains. Ethical principles aim to articulate general values on which everyone can agree and function as practical guidelines.

In the biomedical field , Beauchamp and Childress's four principles intended to cover ethical dilemma (Anderson, Anderson and Armen 2006). Those four principles include (i) **Respect** for autonomy which states that the healthcare professionals should respect the decision-making capacities of autonomous persons; (ii) **Nonmaleficence** requires that the healthcare profession avoid the causation of harm; (iii) **Beneficence** principle states that healthcare profession should promote patient's welfare; (iv) **Justice** states that the patients in similar position should be treated in a similar manner.

Similarly, in policing, they have their own sets of values and principles which defines the exemplary standards of behaviour for everyone who works in policing (College of Policing, 2014). The College of Policing has recently produced ethical principles which includes 9 principles based on the principles of public life. These principles are designed to apply to all public officials delivering public services. These principles include (i) **Accountability**: Individuals must be answerable for their decisions and action; (ii) **Fairness**: To support fairness is the foundation of a justice system. Outcomes should demonstrate impartiality and resist discrimination; (iii) **Honesty**: This principle implies being truthful and trustworthy when making decisions; (iv) **Integrity**: This principle refers to the collection of data and to what extent they are correct to make appropriate judgments about individuals; (v) **Leadership**: One leads by good examples; (vi) **Objectivity**: Decisions made are based on evidence and should be based on best professional judgment; (vii) **Openness**: Individuals are open and transparent in their activities and decisions; (viii) **Respect**: Everyone must be treated with respect; and (ix) **Selflessness**: Individual acts should be focused on public interest.

Each organisation has its ethical code or guidelines they follow. Ethical principles are the basis for an ethical approach towards developing an information and decision support systems. Identifying ethical principles within a particular technology is crucial. These ethical principles vary from technology to technology. Moral values and principles vary from domain to domain. In some domains, accuracy is preferred to other principles. These values

and principles are intended both to motivate morally acceptable practices and to produce ethical, fair and safe machine learning applications.

Over the last few years, there has been some changes to the ethical principles. The US Association for Computing Machinery (ACM) issued a set of seven principles for Algorithmic Transparency and Accountability, outlining guidelines to minimize potential harm when realizing the benefits of algorithmic decision-making (USACM 2017). Around the same time, the Asilomar AI principles developed guidelines on ethics and values that use of AI must respect and consider long term issues (Whittlestone et al 2019). Moreover, around that time several other organisations published additional sets of principles: including Google's 'AI Ethical Principles' (Pichai 2018); Microsoft (Smith and Shum 2018); IBM (Cutler, Pribić and Humphrey 2018). These different sets of principles have considerable overlap. They all convey the message that technologies should be used for the common good, should not be used to harm people or undermine their rights. Based on the different organisation's sets of principles, the common sets of ethical principles that have been discussed by all organisations I have summarised as follows:

- **Accountability**: Developers or people who are involved in designing and deploying a machine learning system must be accountable for how their system operates.
- **Fairness**: Machine learning algorithms should treat everyone in a fair and balanced manner.
- **Transparency**: Machine learning systems that have an impact on individual's lives should be transparent to the people using it for the decision-making process. These systems should be explainable and understandable.
- **Data Privacy**: Machine learning system should be secure and respect the privacy of an individual.

2.7 Visualisation

Machine learning algorithms are to a greater extent providing effective solutions to issues that are related to well-defined tasks such as clustering, classification or predictive modelling based on trend modelling among others (Marsland, 2009).

Drawing defensible conclusions from available information is one of the main tasks of an intelligence analyst. However, the inability of a human to understand the outcome due to algorithmic complexity and making sense of the outcome is becoming increasingly challenging. The field of visualisation, and in particular information visualisation, plays a unique role as an approach to facilitate the human and computer cooperation and makes the results accessible to the end-users. Visualisation can be seen as a mapping from arbitrarily high-dimensional data space to lower dimensions' visual space (Turkay, Laramee and Holzinger 2017) such that the patterns and relations in the data can be more easily perceived. For example, by applying dimensionality reduction techniques, one can project high dimensional data to a 2D visual space as scatterplots that shows clusters and outliers in the data (Sacha, Jentner and et al. 2017; Sacha, Zhang and et al. 2017). A combination of visualisation and machine learning approaches has offered a novel solution for human reasoning for complex algorithms (Endert et al. 2017). Visualisation of machine learning outcomes has become extremely important with the new European General Data Protection Regulation (GDPR), which creates a need for transparency of machine learning outcomes to assess whether a decision conforms to ethical standards (Goodman and Flaxman 2017). The iPCA system, SensePath, ForceSPIRE, and Concept Explorer are good examples of pioneering research efforts in the field.

The interactive PCA (iPCA) system offers different perspectives and interaction within data and model space. This system allows visualisation of the result of PCA using multiple coordinated views. The system allows user to edit or remove data to aid the user in understanding both the PCA process and the datasets (Jeong et al. 2009).

Nguyen et al. (2016) developed a SensePath, a provenance tool, where the analysis process itself can be recorded and visualised to enable browsing through the visual analysis states.

ForceSPIRE is an interactive analysis of textual data within a spatial visualisation. It takes a collection of text documents as input and lays out documents visually such that the layout reflects user notions of similarity and distance. (Endert, Fiaux and North 2012).

Sacha et al. (2017) developed a human-centred visual analytics framework called Concept Explorer to incorporate human knowledge in the machine learning process. This system

allows criminal intelligence to run different machine learning algorithms to visualise crime patterns in an interactive visual interface.

A visualisation tool for Interactive Comparative Case Analysis (CCA) has been developed recently in the field of criminal justice as it has the ability to drive for better decision-making (Sacha et al. 2017). CCA visualisation consists of different exploration areas which helps an analyst to identify groups of crime which consists of similar patterns and key features within that cluster.

Visualisation helps to "make the invisible visible". 'Visualisation' and 'make visible' are two different important concepts that can be used to make:

- the process transparent which shows the use of different types of data including PII and prejudice
- the analysts are aware of the features importance for the accuracy of the algorithms.

Visualisation is a technique that represents information through different forms of diagram, images, animations or any other visual form to denote data. On the other hand, 'make visible' can be considered as a signifier. According to Norman (1986) signifiers are signals. Signifiers can be useful to provide insights into the analysis process allowing the analyst to make ethically aware decisions. These can be in different forms such as signs, labels etc.

The algorithmic decision-making process is complex and consists of different intertwined correlation of decisions. However, to understand the complex system we don't always need a complex solution. Using simple graphs, tables and charts we can untangle and partition the complexity of the algorithmic +process. The visualisation helps:

To improve the transparency process of algorithmic opacity

Machine learning algorithms operate in a black box, where input and output are known, but how or why an algorithm made a certain decision is unknown. Visualisation helps to make the invisible visible. Better use of visualisation and interactivity assists the analyst in communicating with otherwise opaque systems. According to Hearst (2009) visualisation plays an important role to draw the users' attention and make obscure information visible. Visualisation amplifies cognitive abilities to understand the complex process to support a better decision. As stated by Ben Shneiderman, visualisation's potency in revealing the

unusual distribution and interesting cluster can productively encourage people to extend their methods to detect these patterns as well as incorrect, anomalous, inconsistent and missing data (Hullman 2019).

To be accountable for decision-made and comply with EU legislation

An analyst, who is involved in the decision-making process, wants to understand why a certain decision was made by algorithms. When making a decision they need to provide some reasoning as to why they chose a certain decision as a result, they should be accountable for their decision. The decision made by machine learning algorithms has a big ethical and ethically sensitive impact on individuals' lives. An analyst cannot be held accountable if the decision-making process is not known to the analyst. Transparency is identified as a tool to enable accountability. Transparency involves making the process visible to the user. Visibility may relate to the data, algorithms, goal, outcomes, compliance and influence (EPRS, 2019). With the help of the visualisation techniques, a different form of a chart, graphs or any other form of visual display, analysts can understand why a certain decision was made. Also, the new legislation rule "the right to explanation" in GDPR, whereby a user can ask for an explanation of a machine learning algorithm decision that significantly affects them (Goodman and Flaxman 2017).

To understand the outcome quickly and effectively

Too much information can be overwhelming when you are dealing with a large amount of data. By using visualisation techniques, we can make the process visible and open to inspection by colleagues. Visualisation can play an important role in presenting the material in a way that supports the human sense-making and reasoning process (Pohl et al. 2014). As with the saying "picture speaks a thousand words", use of visualisation techniques enables an analyst to view visually presented analytics to grasp any issue related to data, recognise any new patterns or understand the results better. The analyst has to make a quick decision, visualisation techniques present the outcome in a quicker and easier way to interpret. It is easier to visualise a large amount of complex data in visual form rather than attempting to decipher numerous records or reports.

2.8 Ethical and privacy issues in machine learning

In this section, I discuss a number of ethical and privacy issues that arise from the use of machine learning.

2.8.1 Ethical issue: Mosaic effect

Machine learning is becoming an increasingly important medium in the construction of decision procedures. More and more data are becoming available, from which decision can be derived. The use of machine learning algorithm helps in aggregating different pieces of data to reveal complex patterns. These decisions have a real consequence for people and often these consequences are unintended as they are harmful. This happens when private information about an individual is published, or when seemingly innocuous data is mashed and collated with other datasets. To avoid these problems, different techniques such as anonymisation and masking has been used. However, with the increasing sophistication of analytics technique and algorithms, de-identified datasets can be combined with other supposedly anonymous data to re-identify an individual and the data associated with them. Integrating diverse data can spawn entirely new and unknown risk and unexpected consequences, this is known as the mosaic effect (Huber 2014).

The mosaic effect occurs when harmless isolated data are combined and used to paint a clearer and encompassing picture of who an individual is and what he or she likes. The notion of the mosaic effect is derived from the mosaic theory of intelligence gathering. The concept of mosaic effect suggests that even anonymized data, which may seem harmless in isolation, may harm individual privacy if enough datasets containing similar or complementary information are combined. Data that was seemingly innocuous through the eyes of a human, produces insights and predictions that only a machine could infer.

Valentino-DeVries et al. (2018) demonstrated how a person's location data could be pieced together to unveil invasive, detailed personal habits and character traits. A case from Netflix and Internet Movie Database (IMDB) illustrates the potential of the mosaic effect where an anonymous Netflix dataset was de-anonymized by correlating it with the IMDB database. Narayanan and Shmatikov (2008) were able to cross-reference rating scores on rare movies

with those on IMDB which has a public rating to identify the user, uncovering their apparent political preferences and other potentially sensitive information.

In the age of datafication, we leave traces of data in different platforms as part of our daily lives (e.g. posting photos on social media, using a smartphone to navigate around using GPS, using websites). As well as the new development in the field of data analytics, occurrences like the mosaic effect can violate privacy rights and norms and lead to individual harm. Also, analysis can be applied without the knowledge, consent or understanding of the data subjects. Data subjects are the owners of the data. According to Nissenbaum (2004) "*information technology is considered a major threat to privacy because it enables pervasive surveillance, massive databases and lightning-speed distribution of information across the globe*". With the development of new and innovative technologies, the data handling systems have become much more complicated especially in the merging and analysis phase.

The important ethical issue with machine learning algorithms is that people are not aware of how the traces of data they leave behind will be used. S/he has no opportunity to consent or withhold consent for its use. The sheer complexity and opacity of the system make it very difficult to understand, to predict and to control how they behave. As a result, the mosaic effect is capable of affecting an individual's right to self-determination, autonomy and privacy by producing insight and possible inferences into one's private life. The mosaic effect poses a risk involving the transgression of the boundaries put up by a person in his or her private life. The mosaic effect also poses a threat to some important ethical values like privacy and individuality. Also, it can pose a threat to people when personal data is misused or used for a purpose other than what it was collected for. For instance, the Facebook and Cambridge Analytica data scandal involved alleged harvesting and misuse of personal data for an improper purpose (Cadwalladr and Graham-Harrison 2018). Cambridge Analytica harvested the personal data of millions of Facebook users and used it for political purpose without their consent. Moreover, the Cambridge Analytica case shows how mosaic effect allows for highly granular information to be drawn from layering multiple datasets, even when information has been anonymized. The dangers of mosaic effect lie in the different ways in which privacy is threatened.

Privacy is a fundamental human right. Warren and Brandeis (1890) cited by (Raab and Goold 2011) defined the notion of privacy as "*the right to be let alone*" safeguarding the

privacy of individual – freedom from surveillance and the monitoring of behaviour. Each individual has the right not to share certain piece of personal information with a certain group of people. Protecting information privacy (especially when living in the datafication world) of people is an important issue. Informational privacy allows an individual to protect information about him/herself. An individual's privacy can be violated when information concerning an individual is obtained, used or disseminated without the knowledge of the data subject. When information is discovered through a mosaic effect, someone's privacy might be directly violated in the process.

In criminal intelligence, analysts leverage the mosaic effect to obtain a fuller and richer profile of all individual involved, including those who are not under investigation. When the acquired information is classified into profiles and used in decision-making, an individual would feel violated in their privacy. When doing this, the analyst can use the aggregate surveillance data to infer private details about the suspects or individual that no individual member of the public could reasonably learn by observing the suspect for a short time. Schlabach (2014) illustrated an extreme example of Antonie Jones, during a drug investigation, law enforcement officials placed a GPS tracking device to Jones's car. Over the next 28 days, the device tracked the movement of Jones's car. By monitoring an individual for 28 days based on the GPS data, law enforcement learned "*whether he is a weekly churchgoer, a heavy drinker, a regular at the gym, an unfaithful husband, an outpatient receiving medical treatment, an associate of particular individuals or political groups—and not just one such fact about a person, but all such facts*". Through the combination and careful analysis of large datasets, it does bring unique risks that are exacerbated by the mosaic effect.

The combination of an individual, simple and initially unrelated piece of information does provide clearer and encompassing patterns or picture of a person. However, sometimes too much of the information is not relevant. This compromises the privacy of a person. Privacy is a fundamental right of an individual. To protect an individual's privacy and personal or any information against such disproportionate interference in the private lives of individuals, European directives have introduced the principle of **purpose limitation**. The purpose limitation principle states that information cannot be used for other purposes than it was originally collected for. In addition to this, a person whose data is being collected has to be clearly explained the purpose the information was originally retrieved for. The purpose

limitation principle plays an important role in preventing data fusion of databases by separating the information into different silos based on the purpose of collection. Purpose limitation on its own is not sufficient to stop the problem of the mosaic effect. Complexity and opacity nature of machine learning makes it difficult to understand the data fusion and patterns revealed about individuals. However, we need to be aware of the purpose limitation principle. In addition to the purpose limitation, a high degree of transparency is needed to ensure proportionality at all stages of data mining. When building technology, we need to be aware of the legal aspect as well.

Data integrity concerns inaccurate and missing data (Calders and Žliobaitė 2013). This principle applies to the integrity of the data supplied to and provided by the system. Data plays an important role in the decision-making process, therefore, the analyst needs to understand whether the data they are relying on is reliable and the extent to which it is accurate, to make appropriate judgments about it. During the data collection process, data that have been missing or incomplete elements are commonly encountered. However, the important part is how these missing or incomplete data are handled. Ethically it is necessary to provide actual, correct and credible information. Consequently, missing data can create an ethical dilemma.

2.8.2 Ethical issue: Opacity

Paudyal and Wong (2018) describe algorithmic opacity as a condition where the internal workings of computational methods are hidden from the user. There is a growing use of algorithmic systems for decision-making. However, lack of understanding and comprehension of the inner working of machine learning have severe negative consequences affecting an individual as well as a group of people in the society. Algorithmic opacity seems to be one of the main causes leading to an ethical issue in algorithmic decision-making. Algorithms are often highly complex and it is impossible to understand without detailed knowledge or explanation the inner workings of the algorithms especially when they act as gatekeepers and make subjective decisions, this requires ethical scrutiny (Centre for Internet and Human Rights 2015). Machine learning algorithms can tweak operational parameters and decision-making rules (Burrell 2016). As a result, it might introduce uncertainty over why and how a certain decision was made. Understanding the reason behind a particular

problematic decision that might be due to system failure, a one-off bug or bias in the data is almost impossible to understand when the system is opaque (Mittelstadt et al. 2016). The rationale of the algorithm is obscured by the portrayal of machine learning as 'black boxes' and the complex nature inhibits oversight. Pasquale (2015) argued about a growing "black box society" governed by "secret algorithms protected by industrial secrecy, legal protections, obfuscation so that intentional or unintentional discrimination becomes invisible and mitigation becomes impossible." According to Tutt (2016) such problems will increase as algorithms increase in complexity and interact with each other's output to make a decision. Burrell (2016) argues, "*Algorithms are opaque in the sense that if one is a recipient of the output of the algorithm (the classification decision), rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs*." The input and output themselves may be entirely unknown.

Machine learning algorithms are programmed to learn to solve problems. Based on learning they are taught to solve those problems and asked to solve those problems in a different situation (Tutt 2016). Based on learning, machine learning algorithms will make decisions and predictions as human do, but without being able to provide its reasoning. The result obtained can be used by applications where we will be entrusting our fate to the machines we do not understand. There are ample examples of machine-learning algorithms failing in ways we hardly thought about. For example, Tesla remains unsure about the reason for the fatal crash involving its autopilot system. According to Tesla that failure may have been purely technical – the car's radar and camera system may have failed to detect the tractor-trailer that was crossing the roadway or they misidentified the truck as an overpass or overhead road sign (Boudett 2016). HP computers webcam did not track the faces of black people in some common lighting conditions (Sandvig et al. 2016). The web search result was biased towards ethnicity names and found repeated incidences of racial bias (Sweeney 2013). Angwin and Larson (2016) analysed a risk assessment tool that was commonly used in the US criminal justice system. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a risk assessment tool used to predict the likelihood of a defendant recidivism once they are out of jail. COMPAS algorithms were accused of engaging in racial bias, as African-Americans were more likely to be marked as a higher risk of committing a future crime than those of other races. The opacity of algorithms makes it difficult to scrutinize, as a consequence, it leads to lack of clarity to the public in terms of how a certain decision was made (Diakopoulos 2014) and potential incomprehensibility for human reasoning (Danaher

2016). Deconstructing the opacity of an algorithm is not easy. Recently, researchers have focused on providing transparency through visualisation, enhancing interpretability and providing an explanation (Hepenstal et al. 2019).

Burrell (2016) has three distinct types of opacity in algorithmic decision-making:
(i) Intentional Opacity: People who operate and run algorithmic decision-making system do not want other people to know how the algorithms work. The inner working is made opaque deliberately to protect the intellectual property of the algorithms. Furthermore, the main rationale for intentional opacity is to maintain organisation trade secret and competitive advantage over other organisations offering similar services and not allowing people to know how the system or algorithms works, as knowing how the system works allows people with bad intentions to "game the system". Besides, Pasquale (2015) has highlighted three intentional strategies: real secrecy, legal secrecy and obfuscation. Real secrecy establishes a barrier between unauthorized access. Legal secrecy is where they are obliged to keep certain information secret and obfuscation involves making things unclear to prevent tampering and to provide security through obscurity.

If everyone knows how the system works or the different factors which have an impact on the outcome, people will try to skip those factors if they have the chance. Zarsky (2016) in his paper, gave an example on credit score on how a special interest group could try to influence the process to the special factor. For example, if people knew their credit score is calculated based on where they use their credit card and using a credit card in a discount store will make them a bad credit holder. Special interest people will sidestep proxies, which has an impact on the credit score even while still engaging in risk-generating behaviours. For instance, they might stop using their credit card in a discount store or encourage lobbying by the discount owner to remove purchases at these stores from the list of negative factors.

(ii) Illiterate Opacity: This arises due to a large number of people who lack technical skills to understand the underpinning of algorithms and machine learning built from data.

(iii) Intrinsic Opacity: Certain algorithms are hard to interpret due to a fundamental mismatch between how an algorithm and a human understand the problem. In a big data era, with trillions of data, thousands of properties of the data may be analysed. During the analysis process, the internal decision logic of algorithms which generate useful output alters as

algorithms trains themselves on training data. The output produced by the algorithm might be useful and correct. However, the inner logic is not interpretable by humans. The concept of intrinsic opacity can be distinguished into interpretable and non-interpretable as stated by Zarsky (2013). In interpretable, people understand how the statistical output was received, they understand the meaning and can articulate them clearly.

People can indicate the correlations between different features used in the data analysis process. Interpretable can be reduced to human language explanations. Non-interpretable systems are difficult for an analyst to understand as they rely on factors that are too complex for humans to understand. In this process, the software makes its selection decision based upon multiple variables.

Algorithmic transparency is a necessary prerequisite for a democratic society. However, in the current decision-making ecosystem, algorithmic transparency is lagging. The lack of transparency of the decision-making process leaves publics confounded by how and why such a decision was made. Decisions made by algorithms have a crucial impact on an individual's life opportunities - to get a job, release from jail or get caught as criminals and far more. When a decision is made by the algorithmic system, it has a potential impact on an individual's life as it becomes a normative matter to make moral justification and rationale (Citron and Pasquale 2014). Making a system transparent can consist of several steps. Zarsky (2013) in his paper, segmented transparency into three parts:

The collection of data and aggregation of datasets

Transparency refers to providing information regarding the type of information gathered from different information sources such as the forms of data and the databases used in the analysis process. During the aggregation and collation stage, an additional layer of transparency, at this junction, pertains to the human decisions.

Data analysis

This stage includes both technical and human-related aspects. The technology used in the process relates to the technical aspects. Transparency in the technical aspect refers to disclosing information about the software, applications used and, in some cases, if the system is custom-made, transparency can be acquired by releasing a programme's source code. In the United States, there is an Act known as the Data Mining

Reporting Act, which aims to give a "*thorough description of the data-mining technology that is being used or will be used*". According to the Department of Home Security Report (2010), several data-mining ventures use commercial off-the-shelf" (COTS) "*custom-designed software*" provided by private software developers without any additional information about the software. As with the COMPAS algorithm example, they were using software designed by a private company and it appears that a lapse in theory regarding the value and importance of transparency by COMPAS algorithm allowed the practical regulatory practices to sidestep the existing transparency requirements. On the human side, transparency might apply to a variety of elements. The analyst is required to establish a sufficient level of support, which could be acceptable for the "rules" of analysing and confidence, the factors refer to the degree of the rule produced when using automated predictive process.

Usage stage (actual strategies and practices for using predictive models)

Intuitively transparency in the usage stage refers to how the predictive patterns that have been generated are utilized for a particular decision. Furthermore, examining the use of predictive patterns/ models can lead to important insights. It reveals how many of those whom the COMPAS algorithms risk assessment model indicated as a higher risk for recidivism indeed turned out not to re-offend – a false positive. Also, it could further highlight false-negative, that is how many of those considered as lower risk should have been indicated as high risk, yet were missed.

Collection and analysis are two-steps taken to assure data security, retention and tools for access control assure collection and analysis. In the usage stage, transparency pertains to provide data accuracy and to understand if there is an error in the data.

The call for transparency has grown in intensity as more and more organisations and people increasingly use the algorithmic system for decision-making. Transparency helps to understand or know which factors were used or the rate of statistical error in predicting the outcome in any algorithmic processes thus promoting efficiency (Zarsky 2016). Lack of transparency hinders the ability to question the outcome and understand the process of why certain outcomes were derived.

2.8.3 Ethical issue: Accidental discrimination

Accidental discrimination is defined as when the "*outcome of neutral system processes results in the unexpected discrimination and unfair treatment of certain groups or people*" (Marquenie et al. 2017). Everyone seems to agree that discrimination is undesirable and should be eradicated from society. But when using complex technologies, sometimes it is not a person who discriminates, but an automatic algorithm that was programmed without the intention of doing so. We are living in the age of "datafication" where we leave our data-traces everywhere and individual's lives are recorded as "data". Big data does not start as big data but is assembled bit-by-bit from small data and it becomes big when these small pieces of data are compiled into enormous databases (Federal Trade Commission 2013). These data are collected from different mediums and sources. When people provide data for some organisations or when data is collected for a specific reason, people who gave their data rarely have control over how it will be used, aggregated or sold beyond that. A decision made by algorithms is often considered fairer compared to a decision made by a human (Abate and Krakovsky 2018). However, algorithms inevitably make discrimination decisions. When pieces of data are pieced together into big data, the result is densely packed with correlations. As a result, discrimination can occur when using an algorithmic process to inform decision-making.

Discrimination refers specifically to an unjustified distinction or unfair treatment of people based on any physical or cultural trait, such as gender, age, race, religion or sexual orientation (Pedreshi, Ruggieri, and Turini 2008). Discrimination persists in people's lives based on a variety of attributes that are described as protected or Personally Identified Information (PII). Human right laws (Data Protection Commissioner 1995) prohibit discrimination against individual or groups on the basis of race, religions, sex, national origin, age among others. Data-driven decisions can discriminate even in the absence of protected or PII.

Machine learning models are mathematical representations of a real-world process. Machine learning algorithms learn from patterns. To generate machine learning models, we will need to provide training data for the machine learning. For example, police use computational models for classifying risk scores. Given the characteristic of an individual such as location, offense and past crimes, the goal is to predict the score for risk of offending. Based on the

prediction a decision whether “high”, “medium” or “low” risk is made. Using historical data, a model is built. The main objective when building this model is to achieve as good accuracy as possible on unseen new data. Accuracy in machine learning is defined as the fraction of prediction that was predicted correctly by the model. The accuracy, performance and properties of a model depend, among other factors, on the historical data that has been used to train it. Protected attributes such as ethnicity, gender, ages are pervasive, such that machine learning algorithms can learn their correlations when trained on past data.

Recently Dastin (2018) shows a striking example of an algorithm used by Amazon for recruiting people. Amazon’s machine-learning specialist discovered that the recruiting engine they developed were biased against women. This system was trained by observing patterns in resumes submitted to the company over 10-years. The training dataset reflected male dominance across the technology industry. Algorithms learned stereotypical biases tied to gender. Algorithms were prejudiced in making hiring decisions. This example demonstrates how algorithms can learn negative associations for certain protected attributes such as ethnicity, sexual orientation , nationality among others especially when the data reflect a board array of inputs.

Algorithms learning from training datasets have plenty of opportunities to associate protected attributes with statistical regularities, stereotype and past discrimination. A report by White House (2016) argued that the idea of “ hiring for culture fit” introduce biased decisions reproducing perpetuate past hiring patterns. “Unintentional perpetuation and promotion of historical biases, where a feedback loop causes bias in inputs or results of the past to replicate itself in the outputs of an algorithmic system” White House (2016). In the next section, I will discuss the types of discriminations.

Discrimination can be either direct (disparate treatment) or indirect both (disparate impact) of which are protected in recent directives. The following section elaborates on different types of discrimination.

Direct discrimination

Direct discrimination occurs when a person is treated in a less favourable way compared to another person and this difference is based directly on forbidden grounds. For example,

ethnicity discrimination in policing. During the stop and search process, police doing a search based on skin colour or ethnic group. Direct discrimination in machine learning occurs when the algorithm is trained using explicitly membership in protected groups as an input for the model.

Indirect or accidental discrimination

Indirect or accidental discrimination, by contrast occurs when an apparent neutral provision or practice that disadvantages a certain group of people. This discrimination would affect on some people if they have a particular disadvantage compared with others. For example, hiring people with 5 years or more experience for a job might seem neutral at first, but could disproportionally affect those recently graduated or a person who does not have experience for such a long time.

Statistical discrimination is a common term often used in economic modelling (Zliobaite 2015). This refers to inequality between demographic groups occurring even when economic agents are rational and fair. Indirect discrimination is the most relevant type of discrimination that occurs in a machine learning context.

Calders and Žliobaitė (2013) distinguish three means in which machine learning decision-making can lead unintentionally to discrimination (i) data collection for the training data (ii) relationship between attributes in data, (iii) data labelling. These processes are possible sources of discrimination in the training model.

Data collection

The data collection process can be intentionally or unintentionally biased. For instance, according to a report by Gov.uk (2019) there were three stop searches for every 1,000 White people compared to 29 stop searches for every 1,000 Black people. When police conduct stops and searches for a particular ethnic group, even if they were never convicted for conducting any crime or carrying any forbidden items, the historical database will contain skewed data. O'Neil (2016) argues that there is increased police presence in a poor neighbourhood with high crime rates, leading to an increased arrest in those areas. As a result, reinforcing the initial data that led to heavier policing of a poor neighbourhood. When police patrol certain neighbourhood, crime discovered in that neighbourhood will be fed into

the training system, as a result, there is a potential for this sampling bias to be compounded, causing a runaway feedback loop.

Relationship between attributes in data

Some of the attributes are correlated to protected attributes. For example, postcode of a person is highly correlated with ethnicity, as people tend to live close to families or relatives. When the variables are related to each other, it is difficult to identify and control which of the attributes and to what extent they are used in the final prediction. Machine learning algorithms attempt to find correlations in a dataset. For instance, when a company develops a system to help officers decide whether or not a suspect should be kept in custody. The system classifies suspect into "low", "medium" or "high" risk of offending. Those labelled categories are the training data. The algorithm finds which characteristic of the record correlates with being labelled as "high", or "medium" or "low". The set of discovered correlation is commonly called a "model", and these models can be used to predict future outcomes or automate the process of classifying. By exposing machine learning algorithms to decide whether or not a suspect be kept in custody, the algorithm "learns" which are the related attributes or serves as a potential proxy for the outcome of interest. These outcomes of interest in machine learning are defined as "target variable".

Data labelling

Labelling is the process by which the training data is manually assigned by expertise in the field. The true labels in historical data can be objective or subjective. Labels are subjective when human interpretation is involved. In cases of "high", "medium", or "low" for a suspect, labels are subjective, as an analyst who is working in policing and has expert knowledge will label the data. However, when analysts or data miners have to figure out the label for data, it is frequently fraught with peril. Human judgment may influence the subject labelling resulting bias in the target attribute. Machine learning algorithms, which are supposed to remove human bias from the system, often perpetuate unintentional or accidental bias.

When no human interpretation is involved, labels are objective. Such as when classifying if an email is a spam or real, data miners draw from examples that come pre-labelled from when individual people click on spam when they receive the email. Different observers can disagree with the objective labelling; different user may assess a spam email differently.

2.9 Ethical issue: Privacy

When using a system for any processing of personal data, several sources of law have to be considered. Every individual (in legal term: data subject) has a right to decide what information about them can be disclosed to a third party. Traditionally, the right to (informational) privacy is evoked. The right to informational privacy acts as a border against the flow of information and ensures the protection of personal data. Personal data protection is an important aspect of informational privacy. According to European regulation and directives *"protecting the fundamental rights and freedoms of natural persons and in particular their right to the protection of personal right".* However, with rapid technological development, it is questionable whether the right of Data Protection law can adequately protect an individual's personal information. Van den Hoven (2008) highlighted four different moral reasons for protecting personal data. They are (i) Protection against information based-harm; (ii) Protection against informational inequality; (iii) Protection against informational injustice; and (iv)Protection of moral autonomy. Right of informational privacy and data protection law are also relevant in the context of machine learning or AI system.

2.9.1 Data protection law

The Data Protection Act (1998) is a part of legislations, which gives an individual the right to know about the collection, holding, manipulation and distribution of data held about them. The Data Protection Act 1998 deals with the legal framework to regulate the use of 'personal data' in the UK in general. European data protection has its law rooted to OECD principle on privacy protections and the trans-border flow of personal data and the council of European treaty on personal data protection.

The directive has clearly stated the main objective of data protection in Article 1: "*In accordance with this directive, Member States shall protect the fundamental rights and freedoms of natural persons, and in particular their right to privacy with respect to the processing of personal data.*" Laws have been prescribed that deal with the collection and use of personal data that all organisations and individuals need to abide by. Data, personal data

and processing are significant concepts of data protection (Worsley 1987; Woulds 2004). These three concepts are interlinked with each other:

Data

Data encompasses information. Information Commissioner's Office (ICO) has defined data as information processed automatically or manually for a filing system. Under section 1 (1) of the 1998 act, data has been defined as means of information (a) " is being processed by means of equipment operating automatically in response to instructions given for that purpose, (b) is recorded with the intention that it should be processed by means of such equipment, (c) is recorded as part of a relevant filing system or with the intention that it should form part of a relevant filing system." Based on the statement by data protection legislation, it is feasible to partition data into personal and non-personal. Non-personal information does not fall under data protection legislation, so it has not been discussed in legislations. The data protection act aims to safeguard informational privacy by allowing data controllers and regulatory body to process specific types of information which are seen as intrinsically personal or sensitive data.

Sensitive data

Data is considered as sensitive when it consists of information relating to the data subject with regards to racial or ethnic origin, political opinions, religious beliefs or other beliefs of a similar nature, member of a trade union, physical or mental health or condition, sexual life, any offence, or offence committed or alleged to have been committed. Information as such can be used in a discriminatory way so they need to be treated with great care compared to the other personal information (ICO, 2015).

Personal data

Data is considered as personal data when it contains identifiable information about an individual. In article 2 of the directives personal data is defined as " *data which relate to a living individual who can be identified— (a) from those data, or (b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller, and includes any expression of opinion about the individual and any indication of the intentions of the data controller or any other person in respect of the individual;*"

The data protection act sets out the duties and responsibilities of the data controller and the right of data subjects. The act also stipulates eight data protection principles that data controller must comply with when processing personal data. The eight data protection principles that the data controller must abide by are: personal data must be processed fairly and lawfully; collected for specified, explicit and legitimate purposes and not further processed in any manner incompatible with those purposes; adequate, relevant and not excessive in relation to the purposes of processing; accurate and kept up to data; kept for no longer than necessary; processed in accordance to rights of data subjects; kept secure from unauthorised access, unlawful processing, destruction or damage; transferred to a country outside the EU only if that country has an adequate level of data protection. The provision of the Act is translated into different operational documents such as code of practice (ethical principles) to guide UK police.

The focus of this thesis is on Data ethics, Machine learning, and Visualisation. Data ethics involves understanding the use, processing and collection of the data. VALCRI is a system that is designed to assist police in the analysis of large and complex datasets in support of intelligence and investigative work (VALCRI 2014). End-users for VALCRI are from Belgium and the UK. Even though both of them are EU countries and they follow the same EU directives, the Data Protection Law for each country varies. Furthermore, sometimes the concepts are similar but the way each country interprets them is different. In this thesis, I try to understand and compare applicable laws in Germany, Belgium and the UK to determine how legal principles (in the collection, using and protecting of individual data) can be implemented into the system. These include the concept of purpose limitation, ethically sensitive data (privacy information) and the treatment of data subjects.

2.10 Current approaches to solve the issues

This section provides an overview of emerging approaches that are being used to solve the issue related to data ethics, algorithmic opacity, privacy and accidental discrimination. This field of research is relatively new and due to the complex nature of the problem, there is not

one straightforward approach to define the process to solve these issues. Different techniques and approaches have been used.

2.10.1 Legal framework and crowd-sourcing

General Data Protection Regulation (GDPR) is regulation in the European Union (EU) law, which aims to create more consistent data protection and privacy of the individual (Parliament and Council of the European Union, 2016). GDPR specifically focused on the "right of protection of personal data". Recital 71 from GDPR explicit focus on algorithmic discrimination. In the GDPR, algorithmic discrimination is addressed by two key principles: The first principle, data sanitization, which involves the removal of special categories (The European Commission n.d.) from datasets used in automated decision-making. The second principle, algorithmic transparency, introduces the "right to explanation (Goodman and Flaxman 2017), through which data subjects are entitled to "meaningful information about the logic involved, as well as the significance and the envisaged consequences" when automated decision making takes place.
Government bodies, House of Common (2018) are investigating the use of algorithms in decision-making as part of the committee's crowd-sourcing projects.

2.11 Explanation

Explanation in machine learning ensures fairness, identifies potential bias/problems in the outcome and ensures the outcome received meets the ethical standard. To trust the black-box methods, make an ethical decision and to have insight into the causes for a certain decision we need explanation. According to Doshi-Velez and Kim (2017) the explanation can be categorised as local and global. The type of explanation sought depends on whether it aims to explain the entire model or a single decision. A local explanation is built on and around a set of small regions of the trained response functions (Mittelstadt, Russell and Wachter 2019). They can only explain a small section of the response function. In a local explanation, it is easier to decide whether the decision made was using reasonable assumptions. Using local explanation, one would be able to explain the decision even using implicit knowledge. A small section of the response function is more likely to be linear and monotonic therefore, a

local explanation can be more accurate (Lipton 2016). In addition to this, local explanation aligns more with human-scale reasoning and semantic interpretation. In contrast to this, global explanation focuses on how the system works which allows the user to understand the working of a system rather than individual justifications for particular decisions. Global explanation helps us to understand the working of the model (Rieke et al. 2018). The local explanation provides information about the statistical or logical processes involved in a particular case. The local explanation can be more accurate than global explanation, as a global explanation can be approximate based on average values.

2.12 Visualisation techniques

Complex machine learning models are hard to understand. Different visualisation techniques have been designed and developed to help people to understand and explain what models have learned and how they made certain predictions. Some of the Visualisation techniques includes:

2.12.1 Feature importance

Feature importance measures the importance of features by calculating the increase in the model's prediction after permuting the feature. Feature importance provides highly compressed insight into the outcome. Different techniques have been used to measure the important feature within the datasets.

Sensitive analysis

The sensitivity analysis determines what impact each feature has on the model's predictions. Sensitivity analysis inspects whether model behaviour and outputs remain unchanged when data is intentionally perturbed or other changes are simulated in data. This is an approach used by different domains to understand the behaviour of any complex and opaque system. Sensitivity analysis is a model calibration. A complex model system is always dependent upon numerous model parameters. Sensitivity analysis in machine learning allows the relative importance of causative factors in the model to be assessed. The sensitivity analysis approach is often extended to Partial Dependence Plots (DPD) or Individual Conditional

Expectation (ICE). Sensitivity analysis is commonly used to explain feature importance (Montavon, Samek and Müller 2017).

Local Interpretable Model Agnostic Explanation (LIME)

LIME provides insight into feature importance and their linear trend around specific, important observations. LIME can handle non-continuous input features frequently found in machine learning (Ribeiro, Singh and Guestrin 2016).

SHarPley Additive Explanations (SHAP)

SHAP is a unified approach to explain the output of any machine learning. SHAP assigns each feature importance values for a particular predictions. SHAP approach uses Shapley values to quantify the importance of features of a given output. SHAP produces the best possible feature importance explanation with a model agnostic approach (Molnar 2019). Feature importance helps to draw conclusions about features that contributed to the decision making of the algorithm.

Discrimination-Aware Data Mining (DADM)

Berendt and Preibusch (2014) introduce an approach of exploratory DADM (eDADM) to highlight discriminatory information during the analysis process to increase user awareness to that information. eDADM is useful when it is not clear whether a distinction by some attribute amounts to discrimination in the legal sense or not. Providing eDADM allow for more open-ended interpretation and evaluation and, importantly, people using the system will be aware of the discriminatory and non-discriminatory features. This approach encourages keeping human in the decision-loop.

2.13 Gap

The VALCRI system involves the processing of a large amount of data which includes personal and non-personal data. Personal data is at the heart of the EU GDPR, but many

people are still unsure about what exactly is ‘personal data’. In Data Protection law, the term personal data has been defined extremely broadly. According to GDPR (The European Commission n.d.) “‘personal data’ means any information relating to an identified or identifiable natural person (‘data subject’)”. The data that are defined as personal and sensitive data are contextual based. There is no definitive list of what is or is not personal data. It all depends on how people interpret the GDPR definition of personal data. During analysis, data plays an important role, so the selection of data has an impact on the accuracy and also in pattern recognition. Some of the earlier research (Larson et al. 2016) exposed discrimination embedded in data processing software even when personal data was not used during the analysis process. The resulting gaps between the clear guidelines on personal data and opacity in the algorithmic process can have severe consequences affecting individuals, groups, as well as a whole society. Legal framework, visualisation and explanation techniques have been introduced to fill some of the gaps. The different issue raised by algorithmic decision-making is a complex problem and there is no simple solution. As a result, these techniques do not fill all the gaps. Research in this area either focus on ethical prospective or just on privacy aspect. However, ethical and privacy issue should go hand and hand. This thesis looks into the problem from a different viewpoint, to gain a deeper understanding of the problem and provide a solution that helps in making ethical algorithmic decision-making.

2.14 Summary

This chapter has provided an understanding of how machine learning algorithms are being used in the decision-making process. It also discusses the ethical and privacy issue in algorithmic decision-making. It further discusses the current approaches and gap in the literature.

Chapter 3

Comparative analysis of purpose limitation principle across different EU countries

3.1 Introduction

This chapter reports on our study that compares Belgium, Germany and the UK privacy principle in the field of law enforcement. This study was carried out to understand how the term "purpose limitation" is interpreted across selected EU countries (Belgium, Germany and the UK). The aim was to identify commonalities and divergences between selected EU countries. The outcome of the study aims to serve as a basis for understanding the privacy law to safeguard the privacy of individuals when developing the VALCRI system. From our analysis, I found that the UK is more lenient when using personal data compared to Belgium and Germany. In the UK, data collected during the investigation of a specific offense can be shared with other police forces for investigations into other offenses, since this sharing is not 'incompatible' (contradictory) to the purpose for which the data was collected. In contrast to this, in Germany and Belgium, data collected during the investigation of a specific offense cannot be shared with other forces for investigation. Purpose limitation is one of the principles which defines that personal data must be obtained for specified, explicit and legitimate purposes and must not be further processed for any other purpose that is compatible with the original purpose. Nonetheless, while the terminology is the same, due to its flexible and open-ending phrasing, it still differs strongly between the EU Member States.

The principle of the purpose limitation is embedded in all national data protection laws of the European Union Member States, which shows the importance of the principle. The principle of 'purpose limitation' has to be implemented in any technology system which process data including the VALCRI system, since otherwise the processing of the personal data in the system cannot be legally compliant and raises ethical issues such as accidental or indirect discrimination.

In recent years, there is a drastic change in the way data is collected, analysed and applied in the digital era. Instead of determining a problem first, and then looking for data points to solve it, technologies enable an organisation to do the opposite- collect data first, without any specification. Data is collected from multiple platforms as well as through various applications that record users' movements, communications and transactions. The data processing is becoming fully or semi-automated and occurring in real-time.

As more and more personal data become available, the processing capacity has ceased to be a problem. The organisation uses this information to create a profile and to derive insight about people. Supermarkets use profiling to observe actions of the user over time to build a profile for each user, which contains information such as purchasing patterns and shopping patterns. The supermarket uses this data to target advertisements to the user to predict future purchase patterns. These insights may lead to discoveries of people's patterns or behaviour that could violate individual privacy. In one case target sent out ads for baby-related products and diapers to a pregnant girl who lived with her parents, before she even informed her parents of her being pregnant (Hill 2012). Based on her purchasing patterns the supermarket infers the pregnancy of the woman. As Murphy (2017) said "data is the new oil". This has led to companies capturing as much data as possible, then perform analytics on this data to gain valuable insight into the real world. Unlike the previous approach where the problem is identified first, and then based on the problem, relevant data is collected. Recently, the press has highlighted a company, Cambridge Analytica, which was able to obtain data of about 87 million user accounts from Facebook (Cadwalladr and Graham-Harrison 2018). The information was then used to create a profile to target ads, designed to appeal to certain personality types, at people that appeared likely to change their mind about who to vote for in the US presidential elections.

However, one of the fundamental principles for the collection and processing of personal data is to state the purpose of collecting data, which is known as the purpose limitation principle. Purpose limitation is seen as a cornerstone of the data protection regime. The purpose limitation principle aims to prevent the use and reuse of collected data in ways that were not expected by the data subject (the person whose data is being used) (Article 29 Working Party 2013). Defining the purpose of specific data processing is the crucial step towards legal compliance. The principle of purpose limitation is extremely important for fair data

processing. The data controller or the person who collected data must always be able to demonstrate that the data are processed lawfully, fairly and in a transparent manner in relation to the data subject (European Commission 2018).

3.2 Purpose limitation

Article 5 of GDPR lists the purpose limitation principle among other key data principles. It states that personal data must be *"*collected for specified, explicit, and legitimate purposes and not further processed in a way incompatible with those purposes". Specifying the purpose is a crucial first step towards legal compliance. The principle of purpose limitation is designed to safeguard the processing of personal data. The purpose limitation principle has two components:

(i) Collected for 'specified, explicit, and legitimate' purposes and

(ii) Compatible use

When personal data is collected, we usually expect the purpose for which data will be used. According to the Article 12 of the Universal Declaration on Human Rights (Westermann 2018) an individual has the right to decide which information is disclosed to which (third) party and for which reason. This is why purpose limitation is such an important safeguard and cornerstone of data protection. Also, purpose limitation inhibits 'mission creep' that could give rise to the use of the available personal data beyond the purpose they were initially collected for.

On the contrary, once the data has been collected for a specific reason, it can be useful for other purposes which were not initially specified. This principle still allows the data controller to further use the data for other useful purposes as long as they are compatible.

3.3 Importance of purpose limitation principle in systems that process data- The VALCRI system

According to Article 29 Working Party "With the development of multifunctional use of data, it becomes all the more relevant to gain a good understanding of the role and the meaning of the principle of purpose limitation. One of the most dangerous pitfalls would be to reject or weaken the concept simply because its implementation has been too diverse and there is no general understanding of the notion, or because the reality of data processing has changed, and it is a challenge to apply a valid concept to a changed reality. It should be kept in mind that processing of personal data has an impact on individuals' fundamental rights in terms of privacy and data protection. This impacts on the rights of individuals must necessarily be accompanied by a limitation of the use that can be made of the data, and therefore by a limitation of purpose. An erosion of the purpose limitation principle would consequently result in the erosion of all related data protection principles." Similarly, the principle of 'purpose limitation' has to be implemented in the VALCRI system, otherwise, the processing of personal information in the system will not be legally compliant. The end-users for the VALCRI system are from the UK and Belgium and data protection people are from Germany. During our meeting and discussion, I found that purpose limitation is interpreted differently across Europe. VALCRI is a system, which will be used across multiple countries, so it needs to be fit-for-purpose in all countries which means a common acceptable ethical and legal standard.

3.4 Methodology

To understand how VALCRI end-users understood and interpreted the principle purpose limitation, I used qualitative methodology. Data was collected over three phases in this study.

1. Secondary data - First Phase

   I collected some legal documents, which discussed and defined 'purpose limitation' from the EU directives: Belgium, Germany and the UK from different websites. Data collected from the documents were conceptualized through 'coding'. 'Open coding'

was undertaken to identify common theme on the documents. Based on this, four common theme: - Specified, explicit, legitimate purposes, and compatible use was identified. These categories were compared with each other to look for similarities and differences.

2. Observation and meetings- Second Phase

   In this study, observation and meeting were also chosen to gather data from legal partners and end users. This technique was particular suitable because some of the documents were available only in Dutch or French language and some of the information was not available online as some of the European police force do not publish their legal document online. Several observation and meeting took place during VALCRI consortium meeting, work package meeting and meeting between end-users and legal partners.

3. Comparative analysis and Collaboration with legal partners working in the VALCRI

   The aim of coding data in this study was to identify commonalities and divergences on the interpretation of purpose limitation principle. During the comparative analysis, I took the EU's notation of "purpose limitation" as ground truth and compared how three selected nations have interpreted the term. Based on the interpretation a comparative analysis was conducted. Comparative analysis document was distributed among the legal partners for evaluation of the content.

3.5 Results

The principle of purpose limitation continues to be considered as sound and valid. Nevertheless, lack of harmonised interpretation led to differing notations of purpose limitation and incompatible processing in the different Member States (Article 29 Working Party 2013).

3.5.1 The Principle of 'purpose limitation'

Below is the interpretation of 'purpose limitation' by EU and the three member states I investigated.

EU

According to Article 5 of GDPR personal data must be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall, not be considered to be incompatible with the initial purposes."

Belgium

Belgian Privacy Act (DPA, n.d.) , Art. 4: No. 1. states personal data must be:
Collected for specified, explicit and legitimate purposes and taking into account all relevant factors, especially the reasonable expectations of the data subject and the applicable legal and regulatory provisions, must not be further processed in a way incompatible with those purposes.

Germany

Federal Data Protection Act (Bundesamt für Justiz, n.d.), Section 4: (1) The collection, processing and use of personal data shall be admissible only if permitted or prescribed by this Act or any other legal provision or if the data subject has consented. (2) (...) (3) If personal data are collected from the data subject, the controller is to inform him/her as to 1. (...) 2. the purposes of collection, processing or use (...) Part II (data processing by public bodies) Section 14: (1) The storage, modification or use of personal data shall be admissible where it is necessary for the performance of the duties of the controller of the filing system and if it serves the purposes for which the data were collected. If there has been no preceding collection, the data may be modified or used only for the purposes for which they were stored.

UK

Data Protection Act (Information Commissioner's Office, 2015), Schedule 1, Part I 2: Personal data shall be obtained only for one or more specified and lawful purposes and shall not be further processed in any manner incompatible with that purpose or those purposes. Part II (second principle): The purpose or purposes for which personal data are obtained may in particular be specified (a) in a notice given for the purposes of paragraph 2 by the data controller to the data subject or (b) in a notification given to the Commissioner under Part III of this Act. In determining whether any disclosure of personal data is compatible with the purpose or purposes for which the data were obtained, regard is given to the purpose or purposes for which the personal data are intended to be processed by any person to whom they are disclosed

Similarities

The wording used by the different nations is very similar or even identical to one another. The principle of purpose exists in general in all mentioned countries.

The purpose needing to be (i) specified, explicit and legitimate are important key elements and (ii) incompatible use.

Differences

There is quite varied and partially broad interpretation of the term "purpose" in the different countries and differing rules regarding further processing. Incompatibility test partially referring to:

- reasonable expectations of data subject (Belgium)
- balancing test and partially concluding application scenarios (Germany)
- assessment aligned to other principles of transparency, lawfulness and fairness (UK)

3.5.2 Purpose specification

Purpose specification is an essential condition for processing data, as it identifies the aim of the data processing and it lays prerequisite for applying other data quality requirements. As a prerequisite, it will allow a data controller to identify the purpose of each processing. In the

next section, I will define the three key elements- (i) specified, explicit and legitimate purposes within the purpose specification, and I will discuss the similarities and differences among each nation.

Specified

Personal data must be collected for a specified purpose. Purpose specification lies at the core of the privacy principle which aims to protect personal data. To determine if the data processing complies with the law and to understand what data protection safeguards should be applied, it is necessary to identify the purpose of the data collection. The purpose of the collection must be clear and detailed enough to determine what kind of processing is and is not included within the specified purpose.

The data collected must state the purpose for the collection of data. Specifying the purpose of the collection of the data is a key element during the collection of data. The specification for the collection of data has to be given before the collection of data or at least during the time of collection of data. There are many discussions on how precisely the purpose has to be specified.

EU

The purpose must be detailed enough to determine what kind of processing is and is not included within the specified purpose and to allow that compliance with the law can be assessed and data protection safeguards applied. The degree of detail depends on the particular context, the personal data involved and the sensitivity of the data.

Belgium

The purpose must be described in a way that is understandable for everybody, in particular for the average citizen.

Germany

Personal data may only be processed when a given purpose is defined as precise as possible before processing; changing the use of data that is no longer appropriate to the original purpose is basically prohibited (Section 14 BDSG) (Bundesamt für Justiz, n.d).

UK

Police view all information collected (including personal data) as a corporate resource to be shared and linked for the common objective of 'policing' regardless of where the data was collected. The term of ‘policing purposes’ is quite wide and consists of protecting life and property; preserving order; preventing the commissioning of offences; bringing offenders to justice and performing any duty or responsibility arising from common law or statute.

Similarities

Specification has to be prior to the collection of data or at latest at the time the data collection takes place.

Differences

Information Commissioner’s Office, (2015) Art. 29 WP: UK's generalising statement of processing data for ‘policing purposes’ seems too vague and general, as it is impossible to determine which data is processed and how.

Explicit

Personal data must be collected for an explicit purpose. Explicit ensures that the purposes are specified without vagueness and ambiguity. The word “Explicit” has been expressed differently in different countries. In some countries, the terminology focuses on how the purpose has to be explained and in others, it does not necessarily require that the purpose is expressed in any way.

EU

Article 29 (Directives 95/46/EC, n.d.) WP: as much information needs to be expressed and communicated as is necessary to ensure that everyone concerned has the same, unambiguous understanding of the purposes of the processing. The purpose should be specified internally by the data controller and should also be comprehensible for data subjects, data protection authorities and other third parties by notification or similar measures. It shall ensure that the purposes are specified without vagueness and ambiguity.

Belgium

The purpose has to be explained or expressed; as a result, the processing of data for an implicit purpose is illegal.

Germany

Germany has used the term "eindeutig" which means unambiguous instead of explicit. The collection of data should be clear and it should not be more than one interpretation for the collection of data.

UK

UK Data Protection Act does not include the requirement of explicit. However, it requires that the purpose is specified by notification of the data subject or the Information Commissioner.

Similarities

There are no similarities for this key concept as they have different interpretation for the same term.

Differences

All different language versions of the directives are equally binding for the member states. Therefore, the interpretation of 'explicit' has to entail all the different meanings expressed in the different language versions

Legitimate

The specified purpose had to be legitimate. The processing must be at all different stages and at all times based on minimum one of the legal grounds provided in Article 7 (Directives 95/46/EC, n.d.) . According to article 6 (1) (b) the purpose must be per all provisions of applicable data protection law and with applicable laws such as employment law, contract law, consumer protection law, etc. In a broad sense, the requirement of legitimacy means that the purposes must be 'in accordance with the law'.

The specification of a legitimate purpose and lawfulness are two different and cumulative requirements. Article 8 of the EU fundamental Rights Charter distinguishes between the specified purpose and the legitimate basis laid down by law (Directives 95/46/EC, n.d.) . The data processing can only be based on limited legal grounds as outlined in the respective laws.

The purpose should not violate data subject privacy. In addition to this, the processed data must be fair, lawful and must comply with data protection principle.

EU

Article 29 WP (Directives 95/46/EC, n.d.) : the purpose generally has to be in accordance with the law in the broadest sense, including all forms of written and common law, primary and secondary legislation, municipal decrees, judicial precedents, constitutional principles, fundamental rights and other legal principles, as well as jurisprudence, as such 'law' would be interpreted and taken into account by competent courts. Within the confines of law, other elements such as customs, codes of conduct, codes of ethics, contractual arrangements and the general context and facts of the case, may also be considered when determining whether a particular purpose is legitimate.

Belgium

The purpose for processing data must be legitimate. There must be a balance between the controller's and data subject's interest. Purpose that violates data subject's privacy is not considered as a legitimate.

Germany

Germany Data Protection act does not include the requirement of the legitimate purpose.

UK

The purpose has to be in accordance with law basis for each processing. There are three key concepts of legitimate:

(i) A legitimate interest
(ii) A necessity test
(iii) A balance between an individual's interests, rights and freedom

Similarities

Requirement of a legitimate purpose means that the purpose generally has to be in accordance with the law in the broadest sense.

Differences

The UK has provided key concept that is required for a purpose to be legitimate. In contrast to this, Germany has not included the requirement for legitimate purpose. Belgium has given a board guideline for legitimate purpose.

3.5.3 Compatible use- 'not incompatible further processing'

The second component of the concept 'purpose limitation' enables a certain amount of flexibility as it allows further processing of previously collected data, as long as further processing is compatible with the purpose specified at collection. Compatibility emphasises that any further processing of personal data that is not incompatible with the original purpose of the data collection is not against the law. It is useful to first clarify what 'further process' means.

EU

'Processing' means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction; (Article 29 Working Party 2013).

Belgium

"Processing" means any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organisation, storage, adaptation or alteration, retrieval, consultation, use, disclosure by means of transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction of personal data. (Privacy Commission 2009)

Germany

"Processing" means the recording, alteration, transfer, blocking and erasure of personal data." (Treacy et al. 2016)

UK

According to the Association of Chief Police Officers (ACPO) (2006) "'processing' has a very broad meaning, encompassing 'obtaining, recording or holding' the personal data and carrying out various operations in respect of it including organising, adapting, altering, retrieving, consulting, using, disclosing, aligning, combining, erasing or blocking".

Consequently, any processing following collection, if for the purpose initially specified or for any other purposes, of the information constitutes 'further processing' and must meet the requirements of compatibility.

Similarly, the compatibility requirements will be very important in the VALCRI, as most of the data processed will not be collected for this activity. The data protection law does not allow compatible processing, but rather prohibits incompatible processing. This double negation allows some flexibility concerning further use. Also, processing for a different purpose does not automatically mean that it is incompatible. As a result, it needs to be assessed on a case-by-case basis.

Is further processing allowed for other purposes?

EU

Yes, if no incompatible purpose.

Belgium

Yes, if no incompatible purpose, taking into account the reasonable expectations of the data subject and the applicable legal and regulatory provisions.

Germany

Prohibition of further processing with reservation of permission under certain preconditions named by the law.

UK

Yes, if no incompatible purpose.

All countries, accept further processing for other purpose considering
it is not incompatible. There is some specific precondition for further processing, respectively exclusion of incompatibility.

3.5.4 Specific precondition for further processing

Belgium

Further processing of data for historical, statistical or scientific purposes is not considered incompatible.

Germany

Germany has no classical compatibility test. Rather, lists of application scenarios determine the preconditions under which further processing is allowed.

Working Party 29 explicitly have distinguished three categories of further processing of personal data:

Compatibility is prima facie obvious

Data is only processed for the purpose clearly specified or implied at collection further processing may be found compatible. In addition to this, further processing also meets the reasonable expectation of data subjects.

For example, Information that has been gathered for a fraud investigation is disclosed to the criminal court for the subsequent criminal procedures.

Compatibility is not obvious and needs further analysis

There is a connection between the initial purposes and the way data are to be further processed but the compatibility is not obvious. Therefore, assessment is further necessary to determine if additional safeguards are necessary. “The greater the distance between the initial purpose specified at the collection and the

purposes of further use, the more thorough and comprehensive the analysis will have to be" (Article 29 Working Party 2013).

Example: Data collected for the investigation for one crime are to be used for the investigation for another offence.

Incompatibility is obvious

Data are to be further processed in an incompatible way for an additional purpose which is unexpected, inappropriate or objectionable to a reasonable data subject, it is highly likely to be considered incompatible. This would include, personal data of victims of burglaries disclosed to a company for an alarm system for marketing purposes using a secret algorithm, without an appropriate lawful ground and lack of transparency towards the data subjects.

For instance, in 2015 a Professor from Cambridge University created an app called "thisisyourdigitallife" which utilised Facebook's login feature. When people log into the app using Facebook log in, they grant the app's developer a range of information from their Facebook profile, including things such as their name, location, email or friends' list (Cadwalladr & Graham-Harrison 2018). At least 270,000 people used the app using Facebook permitting sharing of personal profile data with the Professor who built the app. The Professor later passed the data to Cambridge Analytica without the knowledge of users. Cambridge Analytica, a data analytics firm, using the data received from Facebook built a "psychological warfare tool". Cambridge Analytica used this tool for major political campaigns – Brexit and US election 2016.

The users were not informed about this processing of their personal data by some third parties. In the context of data protection, transparency is important as it brings trust to consumers by helping them to understand how their data is processed and enable them to challenge the unexpected processing of their data.

The directive provides data controllers with some flexibility with regards to further processing. To assist the data controller in making an assessment of the compatibility of further use of personal data, four key factors for the compatibility assessment have emerged.

These assessments are, nevertheless, open-ended which leaves the assessment susceptible to different interpretations.

(i) The relationship between the purposes for which data have been collected and the purposes for further processing

The greater the gap between the purpose of collection and the further use, the more likely it is incompatible. When further processing is already more or less implied in the initial purpose or the activity is the next logical step after collection, the further processing is likely to be compatible.

(ii) The context in which the data have been collected and the reasonable expectations of the data subjects as to their further use

This focuses on the specific context in which the data were collected and the reasonable expectations of the data subject on the further use of data based on the context. One of the important aspects that have to be considered in this context is the relationship between the data subject and the data controller.

(iii)**The nature of the data and the impact of further processing on the data subjects**

This focuses on the type of data and the impact on the data subject as a result of further processing. Data protection law has been designed to protect individuals against the impact of improper or excessive use of their personal data. The type of data processed plays an important role in all its provisions. As a result, it is important to evaluate whether further processing involves sensitive data or not. As sensitive data is more protected than non-sensitive ones. The more sensitive the information involved, the less likely it is that further processing can be categorised as 'not incompatible'. When evaluating the impact of further processing, both positive and negative consequences should be taken into account. According to Article 29 Working Party (2013) "Again, in general, the more negative or uncertain the impact of further processing might be, the more unlikely it is to be considered as compatible use. The availability of alternative methods to achieve

the objectives pursued by the controller, with less negative impact for the data subject, would certainly have to be a relevant consideration in this context."

(iv) The safeguards applied by the controller to ensure fair processing and to prevent any undue impact on the data subjects

This focuses on providing a data controller with an opportunity to ensure fair processing and to prevent any undue impact on the data subjects. The data controller can compensate for any deficiencies that might be in the first three factors that the original purpose has not been defined clearly enough by implementing technical and organisational measures as one possible remediation tool (such as partial or full anonymization, pseudonymisation, and aggregation of data) to ensure compatibility and to increase transparency offering the possibility to object to further processing (Article 29 Working Party 2013).

3.6 Discussion and conclusions

Purpose limitation is a key principle among data protection principles. This principle is designed to provide boundaries within which personal data collected for a particular purpose may be processed and may be put for further use. This principle of purpose limitation is particularly important in an era where more and more data about an individual are being collected and analysed. This enabled data to be analysed and turned into useful and interesting information that can be used for any purpose.

This principle has a clear link with other principles, in particular to (i) Data minimisation, necessity and proportionality; (ii) Lawfulness; (iii) Transparency, Predictability and Inevitability. According to data minimisation principle (Information Commissioner's Office, n.d.) – data must be adequate, relevant, and not excessive in relation to the purposes for which they are processed and data must be kept in a form which permits identification of data subjects for no longer than it is necessary for the purposes for which the personal data are processed – shows that the necessity of data, as well as the retention period, can only be determined in relation to the purpose of the data processing. While the imprecise notion of

‘not excessive’ led to a diversity of interpretation, it is important to properly specify the purpose in detail to ensure data minimisation.

The lawfulness of data processing and the principle of purpose limitation are two distinct requirements that are important in the context of the VALCRI system. Interference with fundamental rights requires justification on legal grounds.
Therefore, law enforcement agencies are also required to provide legal justification for the processing of data. Nevertheless, while the legal ground might provide a general overall purpose or task, the purpose of a specific data processing has to be specified in more detail.

Transparency is one of the key issues when using machine learning or Artificial Intelligence technology to process data. The process is opaque to the user and data can be used in an illegal way leading to ethical issues. The goal of a ‘specified, explicit purpose’ allows data subject and other stakeholders to understand what personal information is processed and the purpose of data processing lies at the core of transparency.

The purpose of the data processing needs to be specified, explicit and legitimate. This requires that the purpose should be made available in writing or orally, and are so clear that it should not leave any doubt or difficulty to their meaning, scope and methods. The purpose should be specific enough to enable the exposure of the separate processes of personal data processing and to enable the assessment of compliance with laws. Data must not be further processed for an incompatible purpose. The purpose must match the expectation between data subjects and data controllers and allow understanding if any subsequent processing purposes do not correspond to the facts of the case. More importantly, any initial purpose needs to be based on legal ground, be compliant with all applicable laws and legal principles.

The lack of a consistent approach weakens the position of the data subject. This can also impose unnecessary burdens on criminal intelligence analysis operating across borders. In the UK, The Guidance in the Management of Police Information from the Association of Chief Police Officers (ACPO) (2006) indicates that police view all information collected as a corporate resource to be shared and linked for the common objective of policing regardless of where the data was collected. Also, police have the power to process personal data to meet the policing purpose. Policing purpose is quite wide and consists of protecting life and property; preserving order; preventing the commissioning of offences; bringing offenders to

justice and performing any duty or responsibility arising from common law or statute. However, this is not the same in Belgium and Germany. According to section 14 of BDSG (Federal Republic of Germany 2009), personal data may only be processed when a given purpose is defined as precise as possible before processing; changing the use of data that is no longer appropriate to the original purpose is prohibited.

The lack of a consistent approach can impose an unnecessary regulatory burden on policing domain in solving the crime and minimizing the crime rate. Especially as more and more data and their global availability have increased exponentially, and processing of personal data has become an increasingly prominent feature of a technology society.

3.7 Summary

In summary, purpose limitation is an important principle that needs to be considered when processing data to protect the individual's right to privacy. However, due to the very open-ended wording of the purpose limitation, it leaves the concept susceptible to different interpretations. Comparative analysis was done to understand how some countries within EU (Germany, Belgium and the UK) understood the principle of purpose limitation. In this chapter, I found that different nations within EU countries have a different interpretation of the term purpose limitation. Some countries are lenient, and some are very strict about re-use of the collected data.

Chapter 4

Privacy scale

4.1 Introduction

The previous chapter described differences in the interpretation of the term purpose limitation between the UK, Belgium, Germany and the EU. The principle of purpose limitation is key for personal data protection. However, there is a vague notion of personal data “any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly…” leaves room for subjective interpretation. In this chapter, I report a list of information that is considered as personal data (also known as ethically sensitive information) in the domain of the police. I defined ethically sensitive information from two aspects: (a) features that identify an individual, known as personally identifiable information, and (b) sensitive features that discriminate against the individual, known as prejudice information. Research objective two is addressed in this chapter, which is to design and develop a privacy scale that can be used during the analysis process. The goal was to identify different types of information that were considered as ethically sensitive information (features) and assign scores to the identified features depending on the level of their sensitivity. I developed a privacy scale, which consists of a value for each feature depending on the level of their sensitivity. The privacy scale, in this specific context, looks at the features which are highly discriminatory and identifiable information. The proposed privacy scale helps to categorise the information based on their impact level. The designed and developed privacy scale can safeguard individual privacy by making analysts aware of the information that is considered ethically sensitive information. The main focus of this chapter is on the policing data especially data that was used in the VALCRI systems.

4.2 Motivation of the study

Police analysts often need to run complex analysis on their data, for example predictive policing to reduce crime by anticipating criminal activity before it happens (Ferguson 2012)

clustering to detect the crimes pattern (Nath 2006), profiling whether someone is likely to re-offend or not (Angwin et al. 2016) and hotspot analysis to identify where to deploy police resource and assets (Chainey, Tompson and Uhlig 2008). For each of the above-mentioned analyse, typically a subset of attribute values in the data will be selected as input to the machine learning algorithms. These attributes are also called features. Feature selection is important as it not only affects the accuracy of the result but also could introduce ethical issues. The analysis conducted by Larson et al. (2016) found "Black defendants were also twice as likely as white defendants to be misclassified as being a higher risk of violent recidivism. Also, white violent recidivists were 63% more likely to have been misclassified as a low risk of violent recidivism, compared with black violent recidivists." Although there was not any unethical decision on the input features, the result showed offenders were stereotyped based on ethnicity and results were inaccurate. There has been substantial work (Guyon and Elisseeff 2003; Shardlow n.d.; Tang, Alelyani and Liu 2014) on the quality of the features selected, but not on the ethical implication of the feature choice. In addition to this, a lack of transparency raises issues on accuracy, bias and lacks clarification on why an adverse decision was made.

Increasing computational transparency will help individuals understand what information about them is being used during the analysis process. While the importance of the selection of features is recognised, work on the ethical implication of the feature choice area has been limited.

In addition to this, recent evidence suggests that the analytical process is opaque, and the analysts are not aware of the features selected during the analysis process. Transparency of the selection of a feature is hugely important because an individual has the right to know what features about them were being used and how the analysts came up with a certain decision. Opaqueness during the analysis process prevents active participation and comprehension of decision-making procedures (Danaher 2016). Moreover, it is important for analysts to be aware of the different features that could have ethical implication when using them for analysis process. As there are no clear guidelines and open interpretation of the personal data, analysts may have different views on personal data. In the next section, I will discuss information that is considered as personal data.

4.3 Ethically sensitive information

Ethical issues surrounding an individual's data are considered as ethically sensitive information. Individual's data can be classified from two different aspects: personally identifiable information, and prejudice information.

4.3.1 Personally Identifiable Information (PII)

The term "Personally Identifiable Information" is used quite often when discussing privacy and data. Any information that in some way contributes to "individually identifiable" information is defined as personally identifiable information (Narayanan and Shmatikov 2010). The Data Protection Commissioner (1995) uses the term "personal data" instead of "personally identifiable information". Data Protection Commissioner (1995) has defined personal data as "any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or one or more factors specific to his physical, physiological, mental, economic, cultural or social identity".

Some of the information that could identify an individual includes, but is not limited to, social security number, passport number or biometric data. Sometimes linked information about or related to an individual that is logically associated with other information are also considered as PII information such as education, finance or employment information. The PII definition and information that are considered as PII vary from domain to domain. As a result, it is important to identify all PII residing within a specific domain or organisation.

Within the European Union, each country has its legislation based on the adoption of the EU data protection directive, which has its definitions of personally identifiable information. The German Federal Data Protection Act is known as Bundesdatenschutzgesetz, BDSG has defined personal data as "information concerning the personal or material circumstances of an identified or identifiable individual (data subject)" (Bundesministerium der Justiz und für Verbraucherschutz 2014). The terms identifiable and identified are used in a similar way to the EU directives.

Belgium, when adopting the EU directive definition, created a more in-depth and complicated version of what is considered as personal data. Belgium privacy protection divides personally identifiable information into two categories: encoded personal data and non-encoded personal data. Encoded personal data is "personal data that can only be related to an identified or identifiable person by means of a code" and non-encoded personal information refers to "data other than encoded personal data" (Privacy Commission 2009).

The UK has defined personal information as "data which relate to a living individual who can be identified— (a) from those data, or (b) from those data and other information which is in the possession of, or is likely to come into the possession of, the data controller, and includes any expression of opinion about the individual and any indication of the intentions of the data controller or any other person in respect of the individual" (Information Commissioner's Office 2015).

4.3.2 Prejudice information

Prejudice is defined as an unjustifiable attitude towards an individual who is a member of a particular group such as a minority group (McLeod 2008). Prejudice can occur as a result of using sensitive information. Sensitive information is defined as information that is protected against unauthorized access to safeguarding the privacy or security of an individual. Information that is considered as sensitive information has been highlighted by the Data Protection Act (1998) these include ethnicity, race, political opinion, religious beliefs among others. Discrimination is an action that expresses the attitude of prejudice. A study conducted on recruitment agencies by Wood et al. (2009) found discrimination based on candidates' names. According to the report, candidates with white names were 74% more likely to be called for an interview following a job application than candidates with an ethnic minority name despite the candidates having the same qualifications. A similar pattern was observed in the study conducted by Sweeney (2013) which observed advertisements by Google AdSense using racially associated name. The study found "significant discrimination" in advert results depending on the name searched for. It was observed that names typically associated with black people were more likely to produce advertisements related to criminal activity.

It is very difficult to assess, if analysts intentionally used prejudicial information or whether it was a result of a combination of different features or whether when making selections from vast data it was an unintended consequence, which resulted in biased or discriminatory results. For example, looking at the case of the COMPAS algorithm (Angwin et al. 2016) according to the report by ProPublica, they mentioned they had not used any prejudicial information such as race. Sometimes, some of the features that are not considered as prejudice information act as proxies, a feature that acts as a substitute for another, or when different features are combined, they will have the same effect as the prejudice information.

Different information can be considered as PII and prejudice information depending on countries. Some information on its own will not be prejudice or identifiable but when combined they become PII or prejudicial information.

4.4 Proposed privacy scale

The privacy scale was proposed to provide a framework for categorizing features, which are considered ethically sensitive information depending on the level of their sensitivity. For example, information about an individual's medical history or finance account is generally considered as more sensitive than information about an individual's postcode. Similarly, certain combinations of PII data will be considered more sensitive, such as when a name is combined with a credit card number. This is considered more sensitive than using names individually. The different input features used in the privacy scale are the features that have been related to personally identifiable information and prejudice information. For PII the score increases when we combine different features depending on what was added. Just revealing someone's surname will not reveal any personally identifiable information. However, it could be prejudice to a certain ethnicity.

The designed and proposed privacy scale will provide the analysts with:

- A structured approach to feature selection in the context of subject privacy
- An auditable 'privacy score' to algorithms applied to an individual based on features selected
- A guideline to the analyst regarding features that have higher privacy score and should not be used during the analysis process

4.5 Study methodology

This study uses both qualitative and quantitative research methods while developing a privacy scale.

4.5.1 Dataset

For this study, I used the anonymized nominal dataset received from law enforcement, UK. The data consists of a three-year period data in which there were naturally occurring crime cases and criminal networks therein. Nominal datasets comprise a list of all persons who are associated with the crime dataset. Nominal datasets consist of approximately 1.49 million records. The nominal datasets were allowed only for the VALCRI project proposed. Therefore, the dataset is not available to the public. The nominal dataset in particular, was chosen for this study as this dataset consists of information that is related to PII and prejudice data such as ethnicity, surname, address among others. The nominal dataset comprises:

Data	**Description**
crime_Ref	full Crime Number such as 99DE1/000030/00
nominal_ref	Computer generated
surname	Surname or family name
forename	First name(s)
sex	Gender
date_of_birth	Date of birth
role_type	victims (VICT); defendants (DEFE); persons who are probably responsible for an offence (PROB); persons known to be responsible for an offence (RESP); Suspects for committing an offence (SUSP)
street_name	Street name of Nominal's home address. House number is excluded
district_name	District of Nominal's home address
town_name	Town of Nominal's home address.
postcode	Post Code of Nominal's home address.
unidentified_desc	Free text entry. Where named details have not been available or included the fullest description of the Nominal is acceptable.
ea_desc	Ethnicity of Nominal
companyname	Free text entry. In cases of a Corporate responsibility, this allows a company to be

	included as a Nominal.
grid_ref_northing	A six-digit Ordnance Survey Grid Reference.
grid_ref_easting	A six digit Ordnance Survey Grid Reference.
beat	This is LPU (Local Police Unit) and sub area.

Information that was considered as PII or prejudice information were chosen from the list.

4.5.2 Participants

Seven operational intelligence analyst participants from different police forces in the UK and Belgium were chosen as participants. Out of 7 participants, 3 were from the UK law enforcement and 4 were from Belgium (2 each were from Federal Police and Local Police)
I was interested in understanding how analysts within the same domain and between different countries (i) understood personally identified information and prejudice information and (ii) identifying- low, moderate and high – ethically sensitive information for police domain. The analysts have an average experience of 12 years.

4.5.3 Procedure

The study followed a qualitative and quantitative research approach. The design of the privacy scale was informed by literature (Cumbley and Peter 2008; European Commission Justice 2010; European union agency for fundamental rights n.d.; International Labour Office. 1992; McLeod 2008; Solove 2014; Rippert, Weimer and Llp 2009), an interview and a questionnaire with the police analyst. At the beginning, based on the literature, I identified different features that were considered as ethically sensitive information within the nominal data sets (dataset received from officers for the project). The nominal dataset comprises different ethically sensitive information and other relevant information of all the persons associated from the crime data sets.

The first phase of the analysis consisted of interview. The interview was conducted to understand the question related to PII and prejudice. Such as how the terms PII and prejudice term were understood by each analyst, and what information they considered as PII and prejudice information. During the interview, I also verified and validated the identified ethically sensitive information within the nominal datasets with analysts. The nominal

dataset consisted of a few examples of PII or prejudice information. During the interview, I also asked for PII or prejudice information that analysts have to deal with but were not the in nominal datasets.

I had created a Likert scale 1-5; (in which 1 was considered low and 5 was highly ethically sensitive information) based on the nominal data. The analyst had to choose an appropriate value for each piece of information based on their sensitivity and impact level. Based on the value for each information from analyst I proposed the privacy scale table 4-1.

4.6 Results

This section discusses outcomes from the analysed quantitative and qualitative data of the information that are considered as ethically sensitive information. Table 4-1 shows our proposed privacy scale, a weighting scheme for ethically sensitive information based on the level of their sensitivity for the UK and Belgium. This proposed scale was built on values given by analysts based on the sensitivity and impact level of each data.

Table 4-1 Privacy scale for the UK and Belgium

	UK		**Belgium**		**Average**	
Features	PII 1-low , 5-high	Prejudice 1-low , 5-high	PII 1-low , 5-high	Prejudice 1-low , 5-high	PII 1-low , 5-high	Prejudice 1-low , 5-high
Name	1	2	2	2	1.5	2
Surname	1	3	2	2	1.5	2.5
Date Of Birth	2	2	2	2	2	2
Ethnicity	1	3	2	4	1.5	3.5
Gender	1	3	2	3	1.5	3
Town	1	1	2	2	1.5	1.5
District	1	2	2	2	1.5	2
Street	2	1	2	2	2	1.5
Full Postcode	3	2	2	2	2.5	2
First part of postcode	2	1	1	1	1.5	1
Fingerprint	5	2	5	1	5	1.5
DNA	4	1	5	1	4.5	1
CCTV	4	1	2	2	3	1.5
Nationality	1	3	2	3	1.5	3

Based on the privacy scale, I mapped the feature into four different categories. The four quadrants represent the different viewpoints for assessing the level of privacy violation and discrimination.

Quadrant I shows the information that is highly discriminatory and also violates individual privacy. Quadrant II indicates the information that is highly prejudice but does not violate individual privacy. Quadrant III marks an area with no damaging information. The use of this information does not discriminate or violate individual privacy. And finally, quadrant IV, illustrates the information that violates individual privacy but are not discriminatory. The information within each quadrant varies from country to country. Figure 4-1 and 4-2 shows the quadrant matrix of how analysts from Belgium and UK scored the same feature in different level of violation and discrimination.

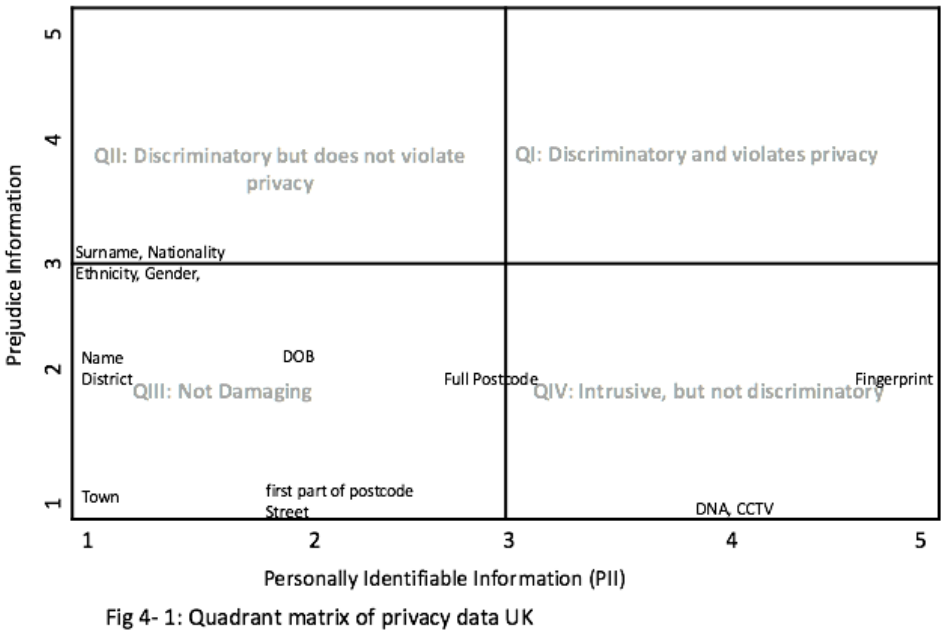


Fig 4- 1: Quadrant matrix of privacy data UK

Figure 4-1. Quadrant matrix of privacy data (UK)

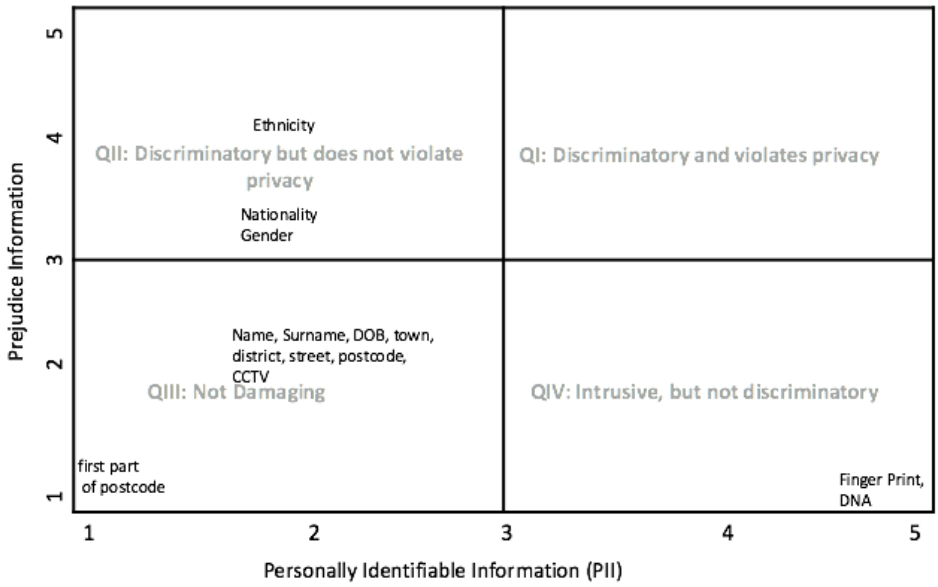


Figure 4-2. Quadrant matrix of privacy data (Belgium)

In addition to the score, analysts were interviewed individually, focusing on information privacy during the data analysis process. Analysts were asked how they define the term "Personally Identifiable Information (PII)". Most of the participants used the term **information** and **identifiable** when explaining the term. However, some of them used an example to define the term. In addition to the definition, analysts were asked to give examples of PII data. Analyst responded with examples such as "Name, surname, address, fingerprint, DNA, partial name/surname, nickname, description of person religion, political preference, sexual and ethnicity, background". The examples given varied from participant to participant. Nevertheless, there was some overlap between the examples. However, participant 2 mentioned '*ear print* in their example'. None of the others mention this example.

When asked how they define prejudice, the response from participant 2 was "*Prejudice is the point when experience turns into the wrong and negative conclusion,"* and participant 3 explained "*the term has negative annotation: making a decision based on PII instead of the data itself. But I think there is a fine line between experience and prejudice"*. Other participants mentioned being discriminatory or biased based on certain attributes. Participants responded with *gender, disability, nationality, ethnicity, religion, educational status, sexual orientation, and surname as* some examples of prejudice information.

According to participant 2 "*a witness describing an offender, or a police report based on what the police officer thinks happened"* are examples of prejudice.
When asked if analysts are flagged or made aware when these PII or prejudice information are used. All participants have the same answer 'no.' Participant 2 mentioned "*No there is no "flagging" there is only a general regulation about all kinds of personal information in data extraction".* Participant 7 mentioned, "*Some are categorized; some are captured unstructured, others not captured is not consistent".* Participant 6 said, *"Categorised - yes but not flagged/highlighted unless it signifies a vulnerability that we need to be aware of".*

4.7 Discussion

The main purpose of this study was to understand and identify different types of information that were considered as ethically sensitive information (features) and assigned scores to the identified features depending on the level of their sensitivity.

The finding suggests that different countries considered different information as personal data. Initially, I used nominal data sets which consist of some of the data that was considered as personal data. In addition to this, during the interview I asked for other information, apart from the list, that they considered as ethically sensitive data.

The result was different than what was expected. From the analysis, it was found that most of the personal information was considered as "not damaging" information by analysts. However, some sensitive information such as ethnicity, nationality and surname were not considered "discriminatory and violates privacy". Nass, Levit, Gostin, & Rule (2009) argued that the data that is considered as private varies among individuals and various groups. Data that is considered intensely private by one person may not be private for other person. Philosophers and legal people would have classified most of the information on list as highly private and some of information such as ethnicity, nationality and surname as highly prejudice information. People will have different interpretation for classifying data to different impact level on individual lives. However, there needs to be a consistency on what data has high impact level. When using machine learning algorithms, data that were

considered as ethically sensitive information can be used due to the correlation of different information.

Earlier research by Royal cited by Solove (2014) showed that sensitive data vary from country to country. In some countries, education is considered sensitive information, whereas in other countries it is not considered as sensitive data. Similarly, the information suggested by analysts such as a nickname, ear print, educational status was considered as personal information by Belgium police, whereas it was not considered as personal information by the UK police. Moreover, I also found that even within the same organisation and the same country, the analyst has a different viewpoint on personal data. Due to the lack of clear guidelines and open interpretation of the personal data, analysts had a different viewpoint for personal data.

During the analysis process, a large amount of data will be collected, some will be personal data. However, apart from the list in the Data Protection Act, it is in the analyst's hand to consider personal information based on their justification. The list of features in the privacy scale will make analysts aware of what features are considered as personal information during the analysis process. Consistency is important during the analysing process. All police analysts will be able to consider the same information as personal information during the analysis process. Furthermore, data plays an important role during the analysis process. Also, all the information that is considered ethically sensitive information or personal information does not expose the same level of harm. Some of the ethically sensitive information can lead to prejudice outcomes but might not have any impact on personally identifiable information. McCallister, Grance and Kent (2010) discussed that each organisation has a unique set of information that are considered as ethically sensitive information. Organisation should apply appropriate safeguard to protect this information based on their impact level.

The purpose of the privacy scale is to provide a guideline for the analyst while choosing the features for Machine Learning. Table 4-1 shows the Privacy scale for the UK and Belgium respectively. These two countries align the same data onto different privacy scales. Belgium considers fingerprint and DNA as highly PII data among others. In contrast to this, the UK

considers fingerprint, DNA and CCTV as highly PII data. In Belgium, ethnicity is considered as highly prejudice data compared to others. However, in the UK, ethnicity and surname falls into the same level of sensitivity. Not all the ethically sensitive information that is available to analysts during the analysis process has an impact on the outcome produced. Analysts should always be careful about the selection of the feature choice. The developed privacy scale provides the analysts with a framework that will help them to make these choices. Figure 4-1 and Figure 4-2 show the different viewpoints on whether it is acceptable to use the data during the analysis process for the UK and Belgium respectively.

The privacy scale is a medium to help the analysts in terms of attribute selections. During the analysis process, the analysts will be aware of what type of ethically sensitive features were used as well as the level of privacy considered. It provides additional description during the analysis process. The proposed scale would allow analysts to understand the level of sensitivity for each feature. Analysts will be able to use these features and be aware of which category they belong to.

4.8 Summary

This chapter identified the differences in understanding ethically sensitive information within the same domain and between different countries. I also found that people working in the same domain and same country see the same information at different impact level. Based on our study, I proposed a privacy scale- weighting scheme for ethically sensitive information based on the level of their sensitivity. The proposed scale has four quadrants, which represent different viewpoints for assessing the level of privacy violation and discrimination.

Chapter 5

Visualising the complexity of computational process using abstraction hierarchy

5.1 Introduction

One of the challenges criminal intelligence analysts face today is handling a large volume of information when solving crime cases. The information is often fragmented, heterogeneous and noisy. Traditional manual approaches are no longer practicable. To solve this problem, Machine Learning (ML) techniques have been introduced into the field of criminal justice to help with data processing and modelling. The ML models are often used to support decision making. For example, based on existing crime data, models can be trained to predict reoccurrences of crime, so resources can be allocated accordingly.

As a decision support tool, the accuracy of ML models is crucial. The ML process consists of different steps that have an impact on the overall accuracy. Moreover, the opaque and complex nature of the ML algorithm makes it difficult for the user to understand and discover the important relationship that runs through the system. Ecologic interface design (EID) has been described as a promising approach to interface design for complex dynamic systems (Reising 2000). In this chapter, I report my findings on the exploration of using the ecological approach for intentional system domains like VALCRI. This chapter describes my attempts at applying the ecological interface design approach that supports decision-making using ML algorithms. Work domain analysis (WDA) is a key step in the ecological approach. Throughout this research, I found that the WDA revealed an important structural relationship between process goals and physical components of the system which could then be visualised as process invariants.

In this chapter, I used VALCRI, as a use case to provide a visual representation of the complexity of the ML algorithm.

5.2 VALCRI: A complex system that integrates machine learning approaches

Visual Analytics for Sense-making in CRiminal Intelligence Analysis (VALCRI) is a system developed for law enforcement. The purpose of VALCRI is to create a visual analytics-based reasoning and sense-making capability for criminal intelligence analysis by developing and integrating several advanced user interface technologies with powerful analytic software (VALCRI, 2014). The VALCRI system uses ML technology to pore over a large amount of data from police records.

One important purpose of machine learning algorithms is to enable analysis of massive quantities of data and enable humans to develop insights into decision-making and predictions. The ML process consists of several steps:

(a) Data pre-processing is an important initial step. When analysing data, it is necessary to make sure that the inputs are suitable for mining. The vast amount of data received by the police is collected from diverse and external sources. As a result, the initial quality of the data will be incomplete (missing values, lacking certain attributes, lacking features values, containing only aggregate data), noisy (duplication of the data, containing errors, outlier values) and contain of inconsistent data. Data preparation involves data cleaning, data integration, data transformation, and data reduction.

(b) Use of appropriate ML techniques. ML is concerned with identifying patterns of characteristics and behaviour based on historical data, which is often used for making a predictive judgment. We need to use appropriate ML techniques depending on what we are trying to solve. Clustering, classification, regression and association are some of the common techniques used in ML. Data collected for analysis will consists of different types of variables such as numerical and categorical. Algorithms such as decision trees can operate directly using categorical data. However, many ML algorithms cannot operate on categorical data directly. As a result, it requires all input variables and output variables to be numerical (Brownlee 2017). This increases dimensionality of the features. Converting categorical values to numerical can cause a model to over fit the

data. Jason (2016) argued that overfitting can lead to poor performance in machine learning.

(c) Data Visualisation is a process that allows the analyst to read and interpret data easily and quickly. Classic visualisation techniques such as histograms, bar charts and scatterplots are effective for small and intermediate size data. However, we face a challenge when we apply classic visualisation techniques to big data due to large amount of data points and dimensions (Tang et al. 2016). Projecting high-dimensional data into space with fewer dimensions is a challenging issue in data mining and machine learning. It is very important to preserve the intrinsic structure of high-dimensional data in the lower dimensional visual display (Sacha et al. 2017).

Although different Dimensionality Reduction (DR) techniques have been developed, the problem of preserving the intrinsic structure of data is not yet fully resolved. Tang et al. (2016) highlight some of issues: (i) performance deteriorates when the dimensionality of the data grows; (ii) sensitivity to different data sets; and (iii) efficiency of the graph visualisation step, which significantly declines when the size of the data increases. Moreover, a study conducted by Paudyal et al. (2017) suggests that depending on the type of algorithm or the features you choose, the result varies. However, some analysts are not aware of these stages or the undesirable consequences they may bring.

More importantly, many analysts are not aware of the fact that the visualisation only approximates the structure of the original data. Some important characteristics of the data may be lost during the DR process. These problems lead to many ethical issues, such as privacy, accuracy, integrity and biased outcome.

VALCRI integrates machine learning to search for semantically similar data across structured and unstructured data in various use cases, such as comparative case analysis, associative search, maps and timeline analysis among others.

VALCRI operates in and exhibits characteristics of complex systems. Complex systems typically demonstrate high numbers of known and hidden interdependencies between components. Outputs from complex systems are so strongly dependent on each other,

changes to system inputs can have unintended, unanticipated consequences. Therefore it is difficult to know exactly which input contributes to an observed output (Ormand 2011). Complex systems exhibit several defining characteristics, such as feedback, strongly interdependent variables and extreme sensitivity to initial conditions. VALCRI has many inter-related and inter-dependent components, such as automated knowledge extraction, text analytics and self-evolving ontologies, based on crime profiles with many emergent outcomes, such as conclusions based on evidence assembled and constructed into explanatory narratives.

Sarter, Woods and Billings (1997) explain that in complex domains, users have to deal with: (i) familiar events; (ii) unfamiliar but anticipated events; and (iii) unfamiliar and unanticipated events (Rasmussen 1985). A major challenge for machine learning in VALCRI is dealing with unfamiliar and unanticipated situations.

5.3 Methodology – case study

Creswell (2013) argues that the choice of qualitative method influences the research study. Qualitative methodologies include narrative research, phenomenology, grounded theory, ethnography and case study. These methods are suitable for a different purpose and are used depending on the research goals and how it is being solved.

The goal of this study is to understand the computational process of the algorithmic system for which a case study was considered to be suitable. A case study is an in-depth examination of one or a few selected instances or events of research interest (McLeod 2019) often involving observation of the system, stakeholders and examination of different sources of information such as text and memos. According to Orum (1992) a case study is an ideal methodology when a holistic, detailed examination is required. Hence, the approach of the case study is to understand the holistic view of the computational approaches of the VALCRI system.

5.3.1 Data collection for analysis purpose

I was able to gather the necessary information required for analysis by attending the VALCRI meetings, looking at the white papers, talking to the VALCRI developers and being part of the team. Working as a team for the project, I had a good understanding of the VALCRI system. Using the work domain analysis approach, I identified the goals and purpose of the system, the priorities the system needs to satisfy, the function and the process. The Part-Whole dimension decomposes the system into a sub-system and physical component. A Part-Whole dimension is an approach within WDA that simply aggregate the physical entities in the VALCRI system at various levels moving up the axis. This decomposition is useful as it shows how the various component of the VALCRI system are organised to perform various functions and processes. I then represented the Abstraction Hierarchy (AH) for the VALCRI system. Abstraction hierarchy is a framework used to analyse complex socio-technical systems, which is commonly used in the field of cognitive engineering (Reising and Sanderson 2002). The visual representation of the abstraction hierarchy is a table of two dimensions. The vertical dimension of the abstraction-decomposition reflects the five levels of abstraction that are coupled in terms of a nesting of means-ends constraints. According to Lintern (2013) the abstract dimension consists of an Abstract Hierarchy that is a diagram constructed through means-ends relations. To represent the VALCRI system in Abstraction Hierarchy, I took one of the purposes – identify and group record according to their similarities- quickly, transparently and in an ethical way- among many other purposes.

AH helped in identifying important functional relationships and system invariants in relation to priorities and value of the VALCRI system.

5.4 Results

(i) Part-Whole dimension decomposes

The Part-Whole dimension decomposes and systematically organises information to provide a big picture of the system. The VALCRI system can be decomposed into three levels: the VALCRI system, subsystems and components. At the VALCRI system level, the system is modelled as a single entity. The subsystems and components represent the detailed granularity of the system. Figure 5-1 shows the part-whole dimension decomposition of the

VALCRI system. There are sixteen components which have been aggregated into six subsystems which make up the overall VALCRI system. DOW, 2014; IEB, 2017; Sacha, Jentner, Zhang, Stoffel, & Ellis, 2017 , explains some of the subsystems of the VALCRI.

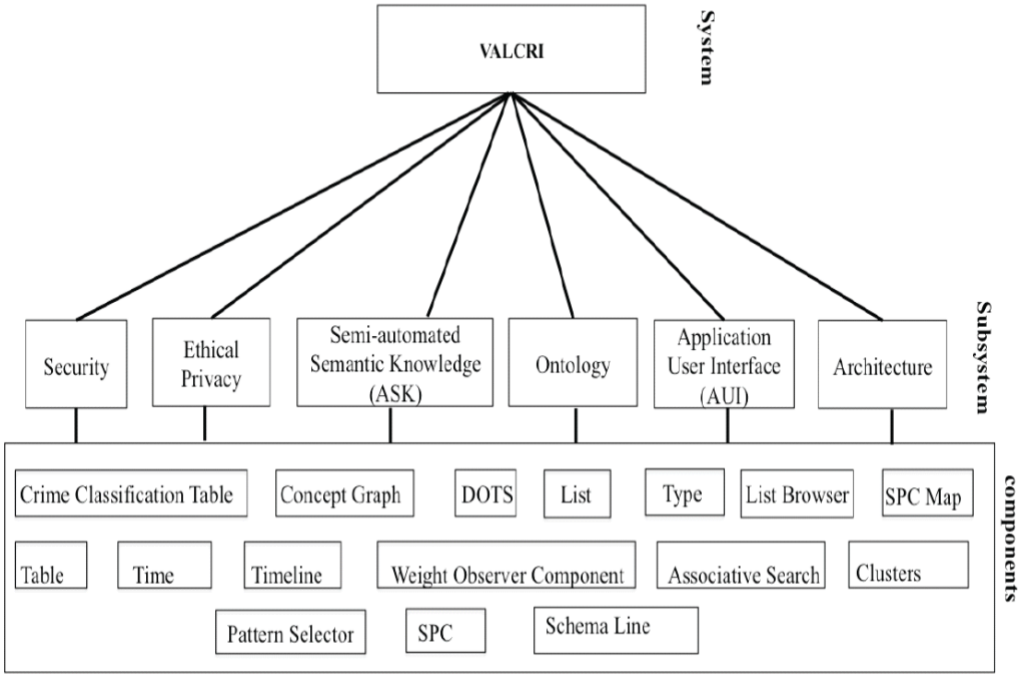


Figure 5-1. Abstraction decomposition of the VALCRI system

(ii) Abstraction Hierarchy (AH)

The AH for our specific use case is represented graphically in figure 5-2 and is outlined in the following sections:

Functional purpose: This level of abstraction corresponds to the rationale behind the design system. One functional purpose (FP) of the VALCRI system can be described as "to group crime reports according to their similarities". One of the important tasks during the investigation is to identify and group crime reports according to their similarity. The similarity of the report is based on the concepts of the features that are chosen. During the analysis process, analysts will receive millions of records and thousands of extracted features from each report. The goal is to identify similarities within the reports.

Abstract function: Usually, abstraction represents the criteria that must be respected for a system to achieve its functional purpose. Criteria are fundamental laws, principles or values, which can serve as a basis for evaluation or judgment. The criteria that must be respected for VALCRI to achieve "group crime report, according to similarities" include ensure quality, few errors, effective data analysis, adhere to the ethical and legal values and finally, aid intelligence analysts in further investigations. An analyst can use the criteria at this level to evaluate how well the purpose-related functions are fulfilling its functional purpose. Abstraction functions allow analysts to reason from first principles. First-principles are important when dealing with unanticipated situations. In the case of VALCRI, they may apply certain heuristics to ensure they are respecting ethical and social values when collecting and processing data.

Purpose- related function: This level represents the function that a system must be capable of supporting so it can satisfy the purpose function. Feature extraction, data preparation, feature selection and dimensional reductions are some of the functions that VALCRI must enable to organise crime reports according to their similarities. Naikar (2013) argues that the purpose-related level can be viewed as describing the "uses" of the object-related functions. Feature extraction points to the uses that selection of features, DR algorithms and their distance calculations, feature correlation, semantic relation between features, etc. serve in the VALCRI. In the VALCRI system, purpose related functions such as feature extraction, data preparation, feature selection and dimensional reductions must be managed in a way that attains "crime report according to their similarities" within the bounds of system's resources.

Object-related function: A system's object-related function serves to archive its purpose-related functions. In the VALCRI system, textual crime report enables the purpose-related function of feature extraction; similarly, the semantic relation between features, calculation and visualisation of feature characteristics, semantic relation between features, transferred to binary vector, etc. enables the purpose related function of feature selection. Object-related functions are highly dependent on the properties of the physical objects.

Physical object: This level represents the physical objects of the system. In the VALCRI system, the representation includes information about each object. The physical object based on the specific use case is Weight Observer Component (WOC) and Similarity Space

Selector (Sacha et.al., 2017). A system's physical object affords a system to achieve its purpose-related function. In the VALCRI system, different type of algorithm are used such as k-means, Principle Component Analysis (PCA) or Multi-Dimensional Scaling (MDS) for dimensional reduction and visual clustering. These are the objects that analysts can change as a consequence the result will vary. Reising (2000) argues that the physical objects represent the properties necessary for classification, identification and configuration for navigation in the system.

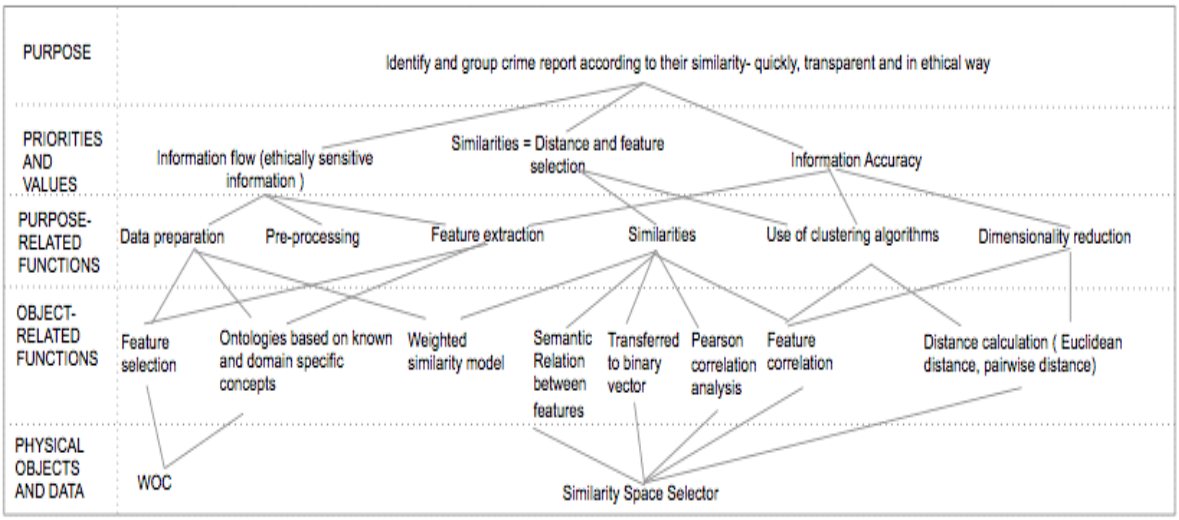


Figure 5-2. Abstraction Hierarchy of the VACLRI system

Our modelling of the AH led us to identify two sets of overarching goals and priorities: (i) enabling analysts to get more accurate outcomes to solve the crime; (ii) adhering to ethical and legal values by respecting information flow (ethically sensitive information) when solving the crime.

Identification of 2 driving functions of the domain

For our work, I focus on the goal of ensuring analysts can archive the highest level of accuracy for identifying and grouping crime reports according to their similarity. This can be decomposed into two priorities that drive the performance of the machine learning algorithms in criminal intelligence analysis: ensure accuracy and ethical information flow. These priorities can be further decomposed to show how the different approaches of machine learning algorithms find a similar crime report for further analysis. We can articulate the relationships between the system purposes, its priorities that drive system performance, the

organisational functions from which we can extract data and the functions that operate the objects of the system.

With the help of the AH, I identified that, to identify and group crime reports according to the similarity, accuracy is important and for accuracy, information flow is also important. The selection of features has an impact on the overall accuracy of identifying crime reports with similar crime types. As a result, the use of ethical information might have an impact on it.

Accuracy

Accuracy is the most relevant function for assessing similar criminal records. Its evaluation can be based on features that were selected during the analysis process. The suggested conceptualization of an accuracy function for similarity entries is

Accuracy= f (feature selection, distance calculation, weighted similarity model, feature correlation, semantic relation between feature, Pearson correlation analysis)

Information flow (ethically sensitive information)

All information including ethically sensitive information has an impact on the outcome of machine learning. Are analysts aware of the impact of this during the data analysis process? The suggested conceptualization of an information flow function for similarity entries is

Information flow (ethically sensitive information) = F (Feature selection, semantic relation between features, feature correlation)

5.5 Semantic mapping of ethically sensitive information

From the AH representation of the VALCRI system, I identified that ethically sensitive information is important during the analysis process. However, ethically sensitive information can fall into different impact levels. Not all the information has an impact on accuracy. In the previous chapter, I identified the different ethically sensitive information and categorised them into (i) Personally Identifiable Information and (ii) Prejudice Information. Based on the sensitivity and impact level, this information is given a score. I

named it privacy scale. Figure 5-3 shows the quadrant mapping for the accuracy and use of ethically sensitive data based on the privacy scale.

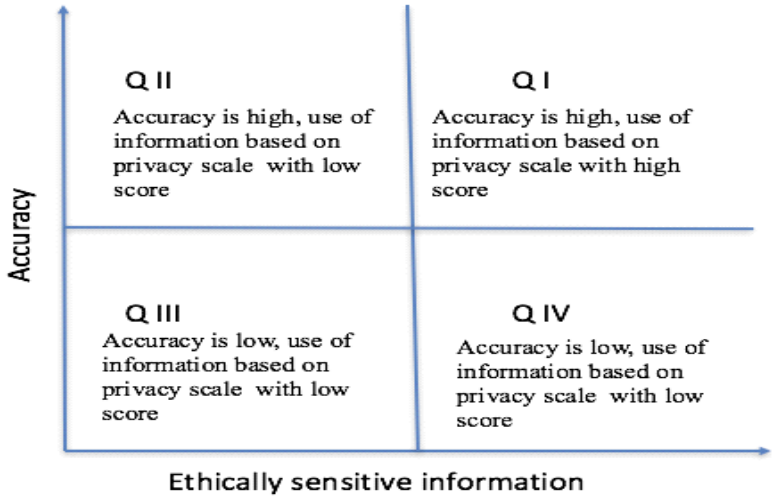


Figure 5-3. Quadrant mapping

The 2- dimensional projection plane is divided into 4 quadrants, displaying different viewpoints for an analyst in assessing the accuracy and use of ethically sensitive data during the analysis process.

Quadrant I: This marks an area of the projection plane, which finds a crime report that has many similarities with a higher accuracy rate, but it uses ethically sensitive information. Using this quadrant, analysts will be aware of the use of ethically sensitive information with high privacy scores. This should be acceptable in the criminal domain as they are using data with highly privacy score to safeguard the public.

Quadrant II: Similarities between crime reports that appears in Quadrant II consists of reports identified with high accuracy but use ethically sensitive information that is classified as low impact.

Quadrant III: Reports projected into this quadrant have low accuracy and also use information that falls into a low privacy scale.

Quadrant IV: This should not be acceptable for analysis purposes. This marks an area where a crime report found has a lower accuracy rate, but it uses information that has a higher privacy impact based on the privacy scale. The use of highly ethically sensitive information doesn't have any impact on accuracy for finding report that are similar.

The quadrant mapping will allow analysts to be aware of the features they are used during the analysis process and they can make choices according to their organisation regulation. Looking at the overall accuracy of the outcome and based on the privacy scale, an analyst can be aware of the consequences and make choices accordingly. This visual representation help analysts to be aware of the privacy information and their sensitivity level. In addition to this, the visualised privacy scale will allow analysts to look at the score for each piece of information during the analysis process to make an informed decision. Table 5-1 shows the privacy scale based on our findings from chapter 4, information with high impact value is highlighted in red.

Table 5-1. Privacy scale

Features	**PII** 1-low, 5-high	**Prejudice** 1-low, 5-high
Name	1.5	2
Surname	1.5	2.5
Date of Birth	2	2
Ethnicity	1.5	**3.5**
Gender	1.5	**3**
Town	1.5	1.5
District	1.5	2
Street	2	1.5
Full Postcode	2.5	2
First part of postcode	1.5	1
Fingerprint	**5**	1.5
DNA	4.5	1
CCTV	**3**	1.5
Nationality	1.5	**3**

In addition to this, one part of a study to determine how features within the privacy scale (finding from the previous chapter) might be important for the analysis process in terms of accuracy but prohibited due to legal reason should be visualised to decision-makers in dynamic intentional domains like criminal intelligence analysis.

5.6 Summary

This chapter describes our attempts at applying the ecological interface design approach that supports decision-making using machine learning algorithms. The approach within the ecological interface design revealed an important structural relationship between process goals and physical component of the system, which could then be visualised as process invariants, such information provided some assistance and awareness to the analysts analysing the criminal reports for more accurate outcomes using machine learning approach.

Chapter 6

Does the use of ethically sensitive information during criminal intelligence analysis process help in better decision-making?

6.1 Introduction

The findings in the previous chapter showed that the accuracy of the outcome depends on the information flow (which includes non-sensitive and ethically sensitive information) through the Machine Learning (ML) process. In this chapter, I investigate the impact of excluding ethically sensitive information in criminal intelligence analysis. The aim is to understand whether and how excluding ethical sensitive information affects the ML outcomes. In order to achieve this goal, I processed a publicly available crime dataset datasets from GitHub (Larson 2016) , and generated two datasets out of it, one with ethical sensitive information, and the other without. The task I choose was to predict whether a criminal will re-offend after being released from prison, which is a commonly used approach by law enforcement agencies. The prediction is often used to allocate resources for crime prevention and reduction. In ML such prediction is achieved by training classification models from existing datasets. Different classification algorithms exist, each have their strengths and weaknesses.

In our experiment, five classification algorithms were selected that are widely used and suitable for the dataset. Same set of algorithms with the same parameter settings were used on the two datasets that was created, one with ethically sensitive information, and one without. the result was compared based on the prediction accuracy and look into the wrong predictions, including false positive cases and false negative ones.

Data used for the analysis process was not balanced and skewed. To understand the result, overall accuracy and per-class accuracy was used. Accuracy is the overall calculation for correctly identified true positive and true negative. In contrast to this, per class accuracy is correctly identified true positive and true negative per class. Analysis shows that using some of the ethically sensitive information during analysis gives more accurate results in predicting criminals who are likely to re-offend once they are out of prison.

In addition, overall accuracy and per-class accuracy were higher using ethically sensitive information. The false-negative calculation was higher for certain ethnicities when using ethically sensitive data compared to without for the different algorithms that were used for the analysis purpose. The outcome was fairer and less discriminatory towards certain ethnicity. Accuracy and fairer outcome are important when relying on machine learning for the analysis process.

Machine learning models are becoming an increasingly important component in decision-making procedures. They are used for the decision-making process that has an impact on individual lives. Recently, the criminal justice fields started using machine-learning approaches in a variety of contexts, including prison rehabilitation programmes, sentencing decisions, informing bail and others. A fairer and unbiased outcome is important when using machine learning for decision making in such a sensitive domain.

Currently, to avoid discrimination and bias in outcomes, ethically sensitive information is removed or suppressed during the analysis process. But, this process does exhibit discriminatory behaviour towards certain groups (Larson et al. 2016; Pedreshi, Ruggieri and Turini 2008). A machine learning method could produce a discriminatory or biased outcome due to the correlation between pieces of information. An individual piece of information on its own may not be sensitive but when combined with other pieces of information, they can reveal information which can lead to a discriminatory outcome. Such as Amazon hiring algorithm which was biased towards women (Dastin 2018), the COMPAS algorithm (Dieterich, Mendoza and Brennan 2016) and O'Neil (2016) in her book "*weapon of maths destruction*" has highlighted many examples.

From our earlier chapter, it was concluded that information flow is important for the machine learning outcome. When we remove ethically sensitive information during the analysis process, we could remove relevant information from the decision-making process. Ethically

sensitive information could be useful information for a fair and accurate outcome. In addition to this, Dwork et al. (2011) argue that to achieve a standard of machine learning fairness one must be aware of how the model's variables have differing predictive power among different ethically sensitive information. To understand if there has been a fair judgment, it is essential to know the value of ethically sensitive information.

According to Cofone (2019) if a machine learning algorithm is "race-blind" and race is removed from the analysis process, then it will be impossible to determine whether the output is discriminatory based on race. As a result, removing ethically sensitive information may not only reduce accuracy but could be self-defeating by reducing the ability to detect bias (Dwork et al. 2011).

6.2 Understanding impact of ethically sensitive information in analysis process

The experiment aimed to understand the impact of including and excluding ethically sensitive information from criminal records when predicting reoffenders. Five different classification algorithms are used to perform classification on the dataset. By experiment, results of all the algorithms will be compared and studied.

6.2.1 Methods

There are different classification algorithms available, including Support vector machine, k nearest Neighbours, weighted voting among others. Classification can be applied to a dataset for discovering a set of models to predict unknown class label. In classification, the dataset is divided into two sets, the training set and a test set. ML algorithms initially run on the training sets, then later the predicting model is applied on the test set to measure the accuracy of the prediction.

6.2.2 Data preparation

ProPublica's COMPAS dataset was used for this analysis. ProPublica obtained datasets (Larson 2016) of defendants and probationers from Broward County, FL. The datasets used

in this experiment contains 53 attributes. From a list of attributes, only 13 relevant attributes and one target value (is_recid) was chosen.. Table 6-1 shows the attribute used for the study with description.

Table 6-1. Crime dataset attributes

Attributes	Data Type	Description
sex	object	Gender of the defendant
age	int64	Age of the defendant
race	object	Race of the defendant
juv_fel_count	int64	Total number of juvenile felony criminal charges
decile_score	int64	Recidivism risk score 1to 10: 1- low, 10-high.
juv_misd_count	int64	Total number of juvenile misdemeanor criminal charges
juv_other_count	int64	Total number of juvenile other criminal charges
priors_count	int64	Total number of non- juvenile criminal charges
days_b_screening_arrest	float64	Days of arrest on screening based
c_jail_in	object	Date defendant went to jail
c_jail_out	object	Date defendant was released
c_charge_degree	object	Numerical indicator of the degree of the charge
c_charge_desc	object	The charge name from Broward
is_recid (Target-variable) -	int64	Numerical indicator of whether the defendant recidivate after 2 years

Different methods are available for attribute selection. For this experiment, the attributes are selected based on their scores of the correlation, with the outcome variable. The ProPublica's COMPAS datasets contain predictions for 7214 records. Out of 7214 records, some of the records contain null values. During our analysis process, records with a null value were removed. For the analysis, 6900 records without any null values were used.

During the data preparation process, decile score was binned to categorical values that represent specific ranges. Decile score contain 1 to 10 value for recidivism risk score. Decile score was binned 1- 3 as low, 4-6 as medium and 7-10 as high. Jail in and jail out were also binned into months.

Some of the datatype within the dataset were categorical as *object* type. Categorical data need to be transformed as some algorithms cannot work with categorical data directly. These data

need to be converted into numerical type. One Hot-Encoding was used to convert categorical data to numerical data (Jason Brownlee, 2017).

6.2.3 Data mining

ProPublica's COMPAS dataset consists of target values. Therefore, the classification approach was chosen for this analysis. Algorithms like k-Nearest Neighbours (KNN) and decision tree are an interpretable model, but they still suffer from an accuracy deficit compared with more sophisticated algorithms like Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting (GB) (Liu et al., 2014; Wodecki et al. 2017). Therefore, a combination of both interpretable and accurate algorithms, including KNN, Random Forest, Gradient Boosting, Gaussian kernel and Ensemble Learning (Max voting classifier) were chosen.

Five commonly used algorithms to demonstrate the variability of the accuracy in the prediction results were selected. Accuracy is defined as the number of correct predictions out of the total number of predictions. True positive (TP) is an outcome where the algorithm correctly predicts the positive class; True Negative (TN) is an outcome where the algorithm correctly predicts the negative class; False Positive (FP) is an outcome where algorithm incorrectly predicts the positive class; False Negative (FN) is an outcome where algorithms incorrectly predicts the negative class (Schwenke and Schering 2007).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Where *TP* = True Positives,
TN = True Negatives,
FP = False Positives,
and *FN* = False Negatives.

6.2.4 Experiment procedure

This analysis was conducted using python. Out of 6900 records used from the ProPublica's COMPAS dataset, classification algorithm was trained on random 80% (5520) of the datasets. According Perner (2011) best results can be obtained by training on 80% of the dataset and testing on 20% of the remaining datasets.

The classification model was trained (i) including ethically sensitive information and (ii) excluding ethically sensitive information. The same training technique was repeated for all the five classification algorithms. Testing was done using 20% (1380 records) of the datasets. The same procedure was followed for the testing datasets. During the analysis process, overall accuracy, accuracy per class, feature importance and the confusion matrix was calculated.

6.3 Results

The accuracy of algorithmic assessment of ProPublica's COMPAS dataset was compared with including and excluding ethically sensitive information based on information from the privacy scale. A positive prediction is one in which a defendant is predicted to recidivate, whereas a negative prediction is one in which they are predicted to not recidivate. Overall accuracy was measured as a number of correct predictions (recidivate or not) out of the total number of predictions (overall accuracy). In addition, false positive (a defendant is predicted to recidivate, but they did not) and false negative (a defendant is predicted to not recidivate but they did) (Confusion matrix) was also measured. Also, the overall algorithmic accuracy, per-class accuracy was calculated (pre-accuracy). When there are few records of one class compared to other classes per-class accuracy will give a very clear picture (Zheng 2015). Records within the ProPublica's COMPAS dataset were highly imbalanced, the distribution of ethnicity, over the 6900 records, is shown in Table 6-2.

Table 6-2. Number of records based on ethnicity

Ethnicity	Records
African-American	3534
Caucasian	2374

Hispanic	584
Other	360
Asian	32
Native American	16

In order to analyse the distribution of race in the ProPublica's COMPAS dataset, a pie chart was used.

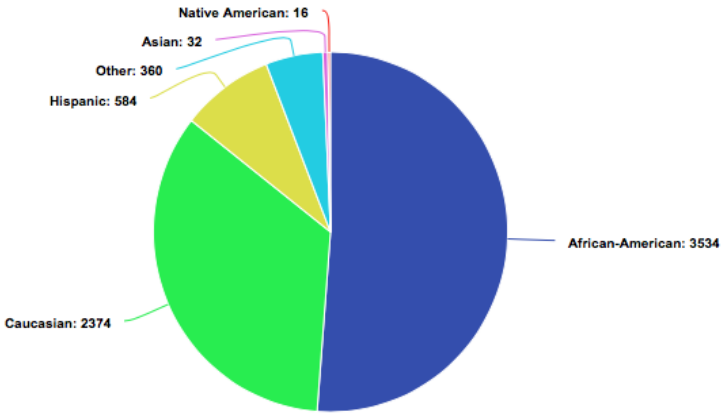


Figure 6-1. Distribution of race in the ProPublica's COMPAS datasets

A simple chart such as pie or histogram can be used to display the distribution of the data. From this chart, we can see the data is highly imbalanced. From the pie chart, it was found that the data were highly skewed to African-American and Caucasian. In machine learning, data imbalance is the most common problem. Imbalanced data has a big impact on the accuracy and bias of the outcome. When we have imbalanced datasets, we can use techniques such as SMOTE, over-sampling and under-sampling techniques (Rahman and Davis 2013). If we didn't use any of these techniques, we need to use different evaluation techniques. The technique that is commonly used in classification evaluation such as ROC (Receiver Operating Characteristic) (Sánchez-Monedero & Dencik 2018), AUC (Area Under Curve) (Weng et al 2017) will not work. In this analysis, per class accuracy was used to evaluate our results.

In this section, results based on (i) Overall accuracy (ii) Confusion matrix based on per-class accuracy will be discussed.

6.3.1 Overall accuracy

Table 6-3. Overall accuracy with and without ethically sensitive information using different algorithms

	k-Nearest Neighbors		Random Forest		Gaussian kernel		Ensemble Learning (Max voting classifier)		Gradient Boosting	
	Using	Without	Using	Without	Using	Without	Using	Without	Using	Without
Classification Accuracy	0.69	0.63	0.71	0.65	0.70	0.64	0.70	0.65	0.71	0.66
False Positive	0.26	0.32	0.27	0.34	0.23	0.42	0.21	0.32	0.26	0.34
False Negative	0.41	0.37	0.37	0.30	0.30	0.37	0.38	0.37	0.32	0.32

Table 6-3 shows the accuracy of the five classification algorithms using ethically sensitive information and without using ethically sensitive information. Classification accuracy has an outcome range between 0 and 1. For classification accuracy a value closer to 1 denotes good accuracy. False positive and false negative are errors in data reporting. In the table 6-3 shows how these errors rate changed with the selection of data.

Classification accuracy gives a high-level understanding of the overall accuracy of the algorithm. However, it hides all the detail we need to diagnose the performance of the model. Based on ProPublica's COMPAS case study, they have two classes (recidivate and didn't recidivate) when calculating the accuracy, both classes are treated equally. Predicting if a defendant will recidivate or not can have a huge impact on individual lives, therefore abstracted information is not enough. Analysts might want to look in detail at how many people recidivate vs didn't recidivate.

6.3.2 Confusion matrix

A confusion matrix provides a more detailed breakdown of correct and incorrect classification for each class.

True Positive (TP) Reality: Person A did recidivate Algorithm: Person A will recidivate	**False Positive (FP)** Reality: Person A did not recidivate Algorithm: Person A will recidivate
False Negative (FN) Reality: Person A did recidivate Algorithm: Person A will not recidivate	**True Negative (TN)** Reality: Person A did not recidivate Algorithm: Person A will not recidivate

Understanding a classification performance is crucial when identifying methods for improvement. This requires a suitable visualisation method. The confusion matrix is a simple and common approach. The confusion matrix in figure 6-2 shows the value for true positive, true negative, false positive and false negative for the Random forest algorithm. The red colour on top right indicates the values for false negative and the red colour on bottom left shows values for false positive. Similarly, the green colour on the left indicates true positive value and on the bottom right shows true negative.

The visualisation of the interactive confusion matrix shows the classifier's correct and incorrect predictions. The X-axis represents the class type predicted by the classifiers and the Y-axis represents the true class type.

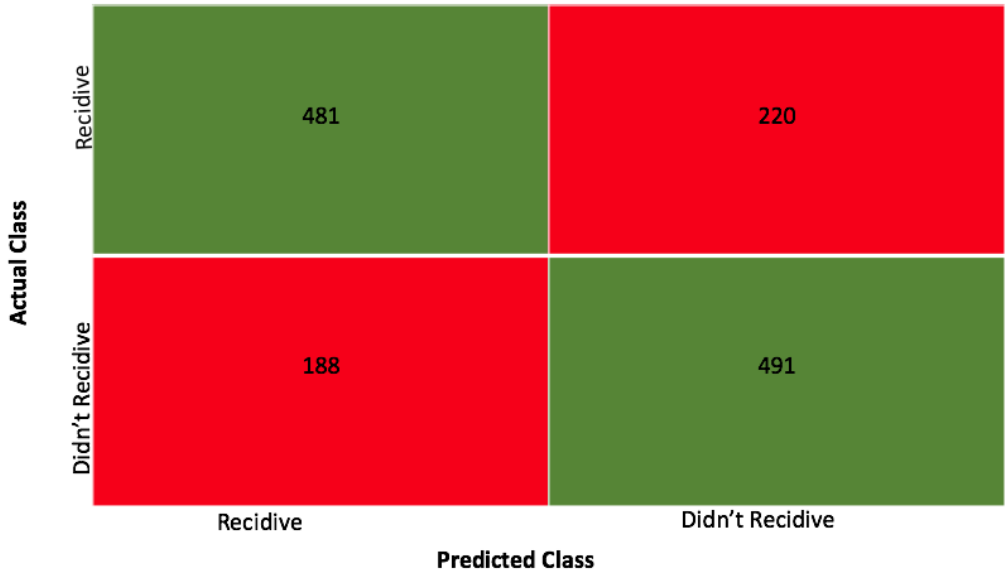


Figure 6-2. Confusion matrix for Random Forest algorithm including ethically sensitive information

By looking at the simple 4 by 4 matrix, the analyst can get useful insight into the errors and more importantly the types of errors that are made during the analysis process. Our confusion matrix is interactive, so if the user clicks on any of the boxes i.e. false positive, false negative, etc., it will elaborate the number based on each ethnicity.

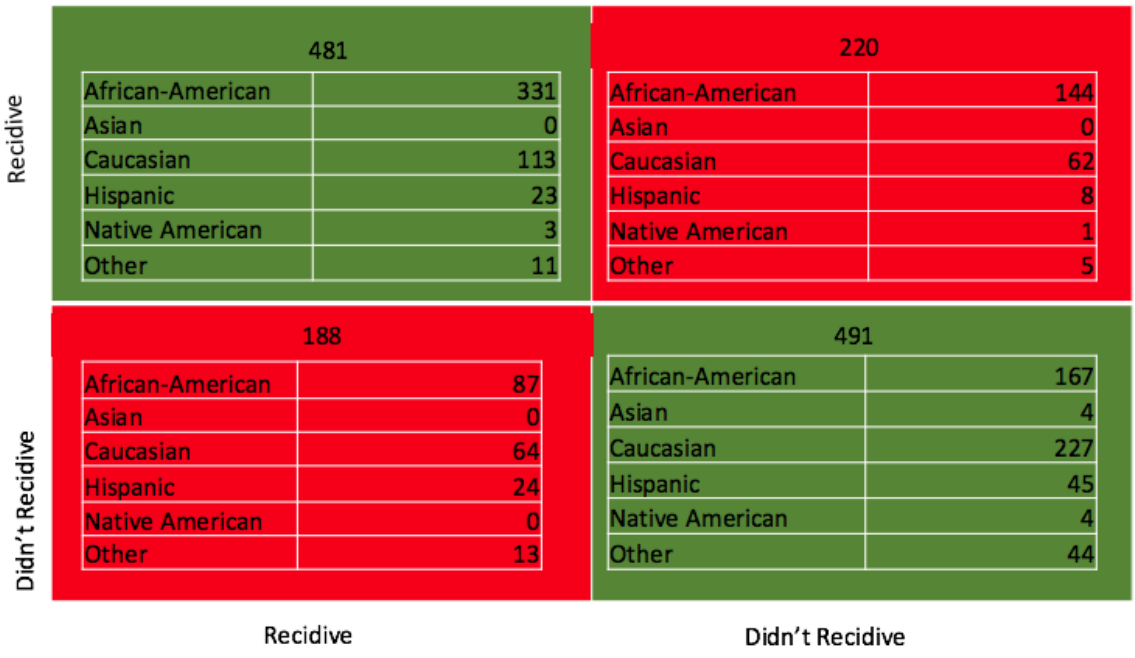


Figure 6-3. Confusion matrix with breakdown of ethnicity for Random Forest algorithm

When user click on any box from figure 6-2, total number of people in each ethnicity will be displayed on the screen as shown on figure 6-3. This helps to understand if the outcome is fair.

Overall accuracy as the rate at which a criminal is correctly predicted to recidivate or not (this was calculated: TP+TN/ (TN+TP+FP+FN)). The result includes false positive (a defendant is predicted to recidivate, but they didn't recidivate) and false negative (a defendant is predicted they will not recidivate, but they did recidivate). In this analysis, the confusion matrix did provide more detailed information for each class. However, as data was highly imbalanced to get a clearer picture for each ethnicity, per-class accuracy evaluation was also measured.

Table 6-4. Accuracy, false negative and false positive mapped with ethnicity using k-Nearest Neighbours (KNN)

KNN	Accuracy		False Negative		False Positive	
	With	Without	With	without	With	without
African -American	67.88%	64.63%	17.53%	17.19%	14.59%	18.18%
Asian	80.00%	37.50%	0.00%	25.00%	20.00%	37.50%
Caucasian	69.67%	65.53%	10.46%	10.00%	19.87%	24.47%
Hispanic	67.96%	59.83%	12.62%	6.84%	19.42%	33.33%
Native-American	100.00%	33.33%	0.00%	33.33%	0.00%	33.33%
Other	68.42%	67.95%	15.79%	14.10%	15.79%	17.95%

Table 6-5. Accuracy, false negative and false positive mapped with ethnicity using Gradient Boosting (GB)

Gradient Boosting	Accuracy		False Negative		False Positive	
	With	Without	With	without	With	without
African-American	69.52%	67.03%	19.94%	16.83%	10.53%	16.14%
Asian	85.50%	100.00%	16.67%	0.00%	0.00%	0.00%
Caucasian	72.28%	70.04%	11.97%	10.79%	15.74%	19.16%
Hispanic	65.44%	62.30%	19.85%	15.57%	14.71%	22.13%

Native-American	83.33%	100.00%	16.67%	0.00%	0.00%	0.00%
Other	79.10%	73.91%	2.99%	7.25%	17.91%	18.84%

Table 6-6. Accuracy, false negative and false positive mapped with ethnicity using Random Forest (RF)

Random Forest	Accuracy		False Negative		False Positive	
	With	Without	With	without	With	without
African-American	68.31%	61.33%	19.75%	17.93%	11.93%	20.74%
Asian	100.00%	75.00%	0.00%	0.00%	0.00%	25.00%
Caucasian	91.01%	71.75%	13.30%	5.15%	13.73%	23.09%
Hispanic	68.00%	58.87%	8.00%	11.35%	24.00%	29.79%
Native-American	87.50%	100.00%	12.50%	0.00%	0.00%	0.00%
Other	75.34%	71.62%	6.85%	6.76%	17.81%	21.62%

Table 6-7. Accuracy, false negative and false positive mapped with ethnicity using Gaussian kernel (GK)

Gaussian Kernel	Accuracy		False Negative		False Positive	
	With	Without	With	without	With	without
African-American	66.23%	66.08%	22.63%	19.30%	13.97%	14.62%
Asian	85.71%	66.67%	0.00%	16.67%	14.29%	16.67%
Caucasian	71.08%	71.63%	9.67%	8.41%	18.24%	19.95%
Hispanic	63.71%	48.44%	13.43%	10.42%	58.96%	41.15%
Native-American	75.00%	80.00%	25.00%	20.00%	0.00%	0.00%
Other	62.87%	55.84%	3.13%	12.99%	37.50%	31.17%

Table 6-8. Accuracy, false negative and false positive mapped with ethnicity using MaxVoting

MaxVoting	Accuracy		False Negative		False Positive	
	With	Without	With	without	With	without
African-American	73.72%	68.38%	14.95%	12.71%	11.33%	18.91%
Asian	88.89%	100.00%	0.00%	0.00%	11.11%	0.00%
Caucasian	69.92%	67.93%	9.20%	7.57%	20.88%	24.50%
Hispanic	66.18%	61.54%	13.97%	8.55%	19.85%	29.91%
Native-American	100.00%	83.33%	0.00%	16.67%	0.00%	0.00%
Other	66.00%	63.33%	20.00%	10.00%	14.00%	26.67%

Table 6-4 to 6-8 show per class accuracy, false negative and false positive for each ethnicity based on five different algorithms. The false-negative value is higher for most of the algorithms when ethically sensitive information was used during the analysis process. Findings indicated that some recidivate cases may be overlooked by the ML models if ethically sensitive information was not included during analysis process. This may lead to miss opportunity of recidivate prevention.

Similarly, the false positive value is higher for all ethnic groups when ethnically sensitive information was excluded during the analysis process. In addition, without using ethically sensitive information offenders who would not recidivate are more likely to be misjudged. This could lead to waste of resources in recidivate prevention. Result does not show any implication that using ethically sensitive information led to biased or discriminatory outcome towards certain ethnicity.

6.4 Overall accuracy

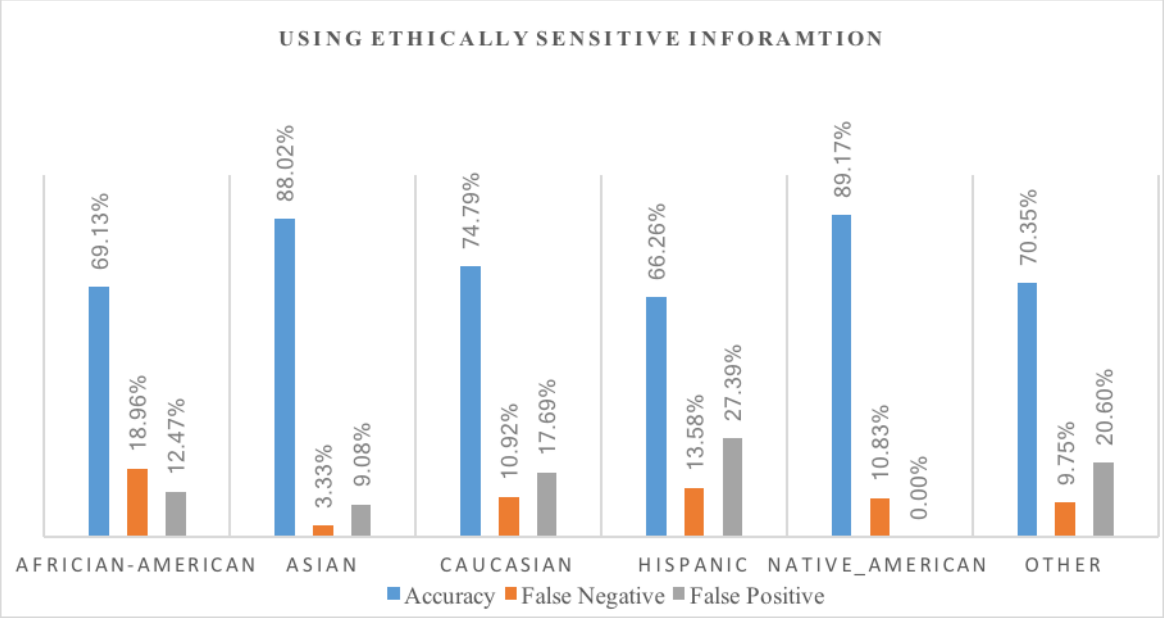


Figure 6-4. Overall accuracy, false positive and false negative score using ethically sensitive information

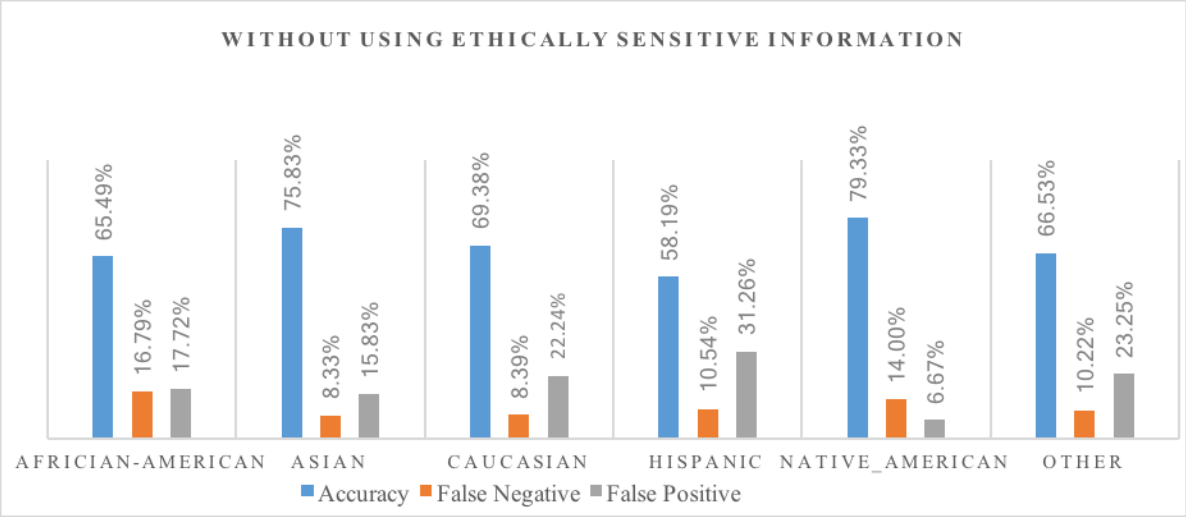


Figure 6-5. Overall accuracy, false positive and false negative score without using ethically sensitive information

Figure 6-1 and 6-2 show the average accuracy, false positive and false negative scores based on five different attributes. The overall average accuracy is higher when using ethically sensitive information compared to without. In addition to this, false-positive scores are lower when using ethically sensitive information compared to without. The values are similar for all ethnicity. However, the false-negative score varies with ethnicity, when comparing including or not including ethically sensitive information. For Native-American, other, Hispanic and Asian the overall false-negative value is higher without using sensitive information compared to with using ethically sensitive information. However, the result is not that much different. This will be discussed more detail in the next section.

Table 6-9. Calculation of per class accuracy for false negative

	False Negative 220	
African-American	144	19.75%
Asian	0	0.00%
Caucasian	62	13.30%
Hispanic	8	8.00%
Native American	1	12.50%
Other	5	6.85%

Table 6-9 shows the percentage of false-negative for each ethnicity. From the table, it can be seen that all values are high for all ethnicity except for Asian. This is due to the lower number of records for the Asian ethnicity. Based on this, it can be confirmed that the result was fair towards all ethnic groups. Some of the ethnic groups have a lower value, this is due to the low number of records for those ethnic groups. However, if our percentage was lower for African American or Caucasian, this denotes discriminatory or biased results. False-negative indicates a defendant is predicted not to recidivate but they did. Visualisation of the breakdown of ethnicity per class helps the analyst to judge if the result is ethical. If they spotted low false-negative numbers for certain ethnicity which have a higher number of records in the dataset, they can rerun the algorithms or change the feature. This allows analysts to be aware of this type of discriminatory outcome.

6.5 Discussion

The results from the study provides some evidence that using ethically sensitive information during analysis process do impact prediction for re-offending. It was found that the use of ethically sensitive information does have impact on correct prediction for re-offending. Using the ethically sensitive information I was able to achieve 6% more accurate results (on average). In addition to this, the outcome was fair to all racial groups. During the analysis process, analysts have to adhere different ethical principles to make decisions that are fair, accurate, transparent, and provide justice. We have seen amplified discussion based on biased and discriminatory outcomes. As machine learning systems develop in complexity, it is becoming increasingly hard to ensure all the ethical principles are being adhered to during the analysis process. Therefore, it is important to identify and assess trade-offs. An appropriate balance between competing ethical principles depends on the specific sectoral and social context an organisation operates in, and the impact on data subjects. The use of machine learning algorithms can lead to biased or discriminatory outcomes. In solving problems that pay attention to ethically sensitive information, trade-off between accuracy and fairness has to be considered.

Accuracy and fair outcome are both important during an investigation process. Biases may manifest themselves in the form of higher false-positive for certain ethnicity that are more

likely to be falsely considered as “high risk”. In the case of the analysis done by ProPublica, it was found a higher false positive rate and lower false-negative rates for black defendants compared to white defendants. The algorithm mistakenly identified black individuals as high risk twice as much as it did for white individuals and mistakenly identified white individuals as low risk more than it did black individuals. The likelihood of a non-recidivating black defendant being assessed as high risk is nearly twice that of the white defendants.

In this experiment, false positive and false negative scores are similar for all the groups. False-positive scores were lower when using ethically sensitive information compared to without and the scores were similar for all ethnicity. False-negative were higher for all ethnicity except in one or two algorithms, where false negative was lower for certain ethnicity. But overall, false-negative was higher for all the ethnicity when including ethically sensitive information.

In some countries or regions, data protection laws prohibit the use of certain ethically sensitive information to protect the individual from biased and discriminatory results from machine learning. However, from this study, it was found that the use of these features contribute towards a more accurate and fairer outcome. During the analysis process, the result was fairer and more accurate when ethically sensitive information was included compared to when ethically sensitive information was excluded. The results indicates that when excluding ethically sensitive information, the ML result is less accurate and even tend to be biased towards certain ethnicity.

In a sensitive domain such as criminal justice, accuracy is important. Therefore, trade-offs have to be made between improving accuracy and respecting privacy based on the ethically sensitive information. In addition to this, the ethically sensitive information’s impact depends on what information are we using (chapter 4 privacy scale). Not all the ethically sensitive information is required for a more accurate and fairer outcome. For ML to be fair to all ethnicities, the false-negative and false-positive value should be the same for all the ethnic groups, it should not be high for one ethnicity and low for others.

The accuracy rates for different algorithms vary. As Breiman (2001) cited by Hall & Gill (2018) stated that for the same input and same target variable different algorithms will produce a very similar but different result. Out of the five different algorithms, random forest

and gradient boosting algorithms gave better accuracy when comparing using and not using ethically sensitive information during the analysis process.

6.5.1 Feature importance

A bar chart displaying feature importance is presented to the analyst and this was produced using the Random forest algorithm to demonstrate a method for proving important feature information. This technique allows the transparency of important and crucial information such as which features had an important role in overall performance. Ethically sensitive information was used for the analysis purpose, to check if the use of ethically sensitive information has any impact of that specific analysis. The analyst is aware that they have used ethically sensitive information for the analysis purpose.

Feature importance visualisation shows the feature that has an impact on the overall performance. Feature importance provides a highly compressed, global insight into the model's behaviour (Molnar 2019). According to Albon (2017) feature importance makes the algorithm simpler to interpret. Figure 6.6 shows the feature importance in a bar chart. Feature importance is shown in two different colours. The red bar shows ethically sensitive information, whereas the grey bar shows not- sensitive information. By looking at the feature importance bar, the analyst can understand which ethically sensitive features have been used for a particular analysis process. There was many ethically sensitive pieces of information used, but the information displayed in figure 6-6 had an impact on the overall performance. Age, Gender and Ethnicity were some of the important features which were part of ethically sensitive information.

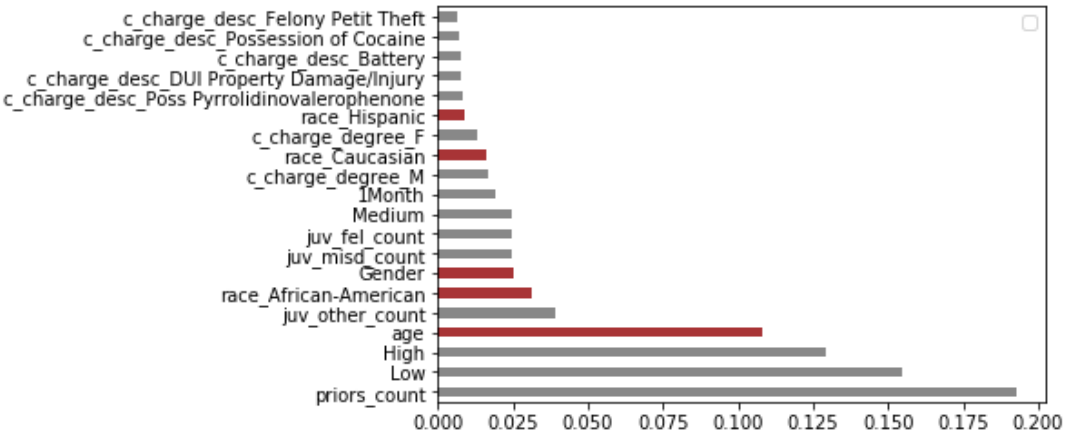


Figure 6-6. Feature importance based on Random Forest algorithm

It was found that ethically sensitive information does have an impact on the result. However, not all of the information is important. Due to different interpretations of the term and no clear example of ethically sensitive information which includes, “personal identified information” and “prejudice” information, it is difficult for analysts to be aware of this information during analysis. Typically, while solving a problem, the analyst is unaware of the ethical implication of these features. Visualisation of feature importance with ethically sensitive information makes analysts aware of this information during analysis. Depending on the organisation’s policy and also depending on how accurate the result improves by using the information, the analyst can decide whether to use or not use the information. Making features visible allows for more open-ended evaluation. It also makes people using machine-learning algorithms aware of the complexity of the notion of discrimination.

6.5.2 Privacy scale

The privacy scale was visualised on tabular format to show analysts different features that are considered as ethically sensitive information. The score for each attribute based on PII and prejudice is also displayed next to the feature to make analysts aware of these scores.
Figure 6.7 shows the visualisation of the privacy scale.

Privacy Scale			
	Feature	PII	Prejudice
1	Name	2	2
2	Surname	2	3
3	Date of Birth	2	2
4	Town	1	2
5	District	2	2
6	Street	2	2
7	Ethnicity	1	4
8	Full Postcode	2	2
9	First Part of Postcode	1	1
10	Finger Print	5	2
11	DNA	5	1
12	Gender	2	3
13	CCTV	3	2
12	Nationality	2	3

Figure 6-7. Privacy scale

Fig 6-7 highlights ethically sensitive information that was used during analysis process. Any information with score 4 or more score either on PII or prejudice based on privacy scale is shown on red. In contrast, any information that has score less than 4 based on privacy scale, is shown on orange colour. So the analyst is aware of these features. Any information that is dark red indicates a high impact on ethically sensitive information. Orange indicates it is

ethically sensitive information with score less than 4. Based on the score they can decide to use or not to use it.

It should be noted that the ProPublica's COMPAS data was highly imbalanced, as it contains a much higher number of records for African-Americans compared to other ethnicities. 51.22% of the records were for African-American, 34.41% for Caucasian, 8.46% for Hispanic, 5.22% others, 0.46% for Asian and 0.23% for Native Americans. To understand the distribution and impact on each ethnicity, per class accuracy for the confusion matrix was used. More importantly, the breakdown of a confusion matrix based on each ethnicity provides analysts with insightful information. Using ethically sensitive information, I was able to correctly predict 69.13% of the records for African-Americans compared to 65.49% without using ethically sensitive information. Similarly, for most of the remaining ethnicity, the result was more accurate when used ethically sensitive information during the analysis process.

6.6 Summary

This chapter presented a study to understand the impact of including ethically sensitive information during the analysis process. To test the impact, an analysis was performed to compare the accuracy of the analysis using ethically sensitive information against the analysis without using ethically sensitive information. Findings suggest that the use of ethically sensitive information has an impact on the accuracy and fairness of the outcome of correctly predicting someone's recidivism once they are out of prison.

Chapter 7
Conclusion

7.1 Introduction

This chapter concludes the main finding of this research and presents a discussion of the results for each research question. It draws conclusion from the results, discusses issues and explores the significance of the research as well as the limitations. Some further research is also suggested as possible extensions of this study.

7.2 Overview of the research

With the growth of big data, machine learning has become an important key to aid decision-making. The main purpose of this research is to understand the ethical impact of the algorithmic decision-making process to support analysts to make ethical decisions. The need for such an undertaking arose from the growth in the use of algorithmic decision-making in application domains such as healthcare, criminal justice, education, recruitment and other fields which have ethical consequences. It was ascertained from the literature reviewed in chapter 2 that lack of ethical awareness in algorithmic decision-making leads to various ethical issues.

Firstly, the black-box nature of the algorithmic decision-making process hinders the visibility and ability to question and understand the process. Secondly, data is an important asset in machine learning. However, there is a lack of clear guidelines on what "privacy information is" as different people will consider different information as private or ethically sensitive within the same domain. Thirdly, when building a system that will be used by different countries, we need to respect the privacy laws. Conversely, when different countries have different interpretations of the same privacy principles, it can create a problem.

The research began with a broad research question "*How can the discrimination- aware approach support analysts for making an ethical decision in criminal intelligence?"* Followed on from that, a literature review was carried out to develop an understanding of the occurrence of ethical issues when using algorithmic decision-making process. The ethical issues raised by the Independent Ethical Board (IEB) of the VALCRI project were also discussed in the literature review. All the information gathered were used during the process of designing and developing of the privacy scale and visualisation techniques that could support the analyst for discrimination awareness during decision making. This research focused on data ethics, machine learning and visualisation to provide a solution for discrimination awareness during decision making in criminal intelligence.

A mixed methodology was carried out to understand and examine ethical issues. An interview was carried out to understand data ethics issues- how different countries interpreted the same privacy law differently; how people within the same domain considered ethically sensitive information. An experiment was conducted to understand the impact of data ethics in machine learning and visualisation techniques were used to provide transparency for analysts who use machine learning algorithms for decision-making.

7.3 Approach to the research questions

The purpose of this research was to support the analyst for ethical decision-making in a machine learning environment using a discrimination-aware approach. The ultimate goal of this study is to investigate different types of ethical issues that could occur in big data analytics and to seek solutions that help improve the transparency and understanding of the data and analysis process. The privacy scale and visualisation tool proposed in this study benefits the analyst. The objectives of the study include:

- To identify the requirements for compliant data processing according to the EU data protection law and how is it interpreted in different EU countries (Belgium, Germany and the UK)?

- To identify the data within the police domain, that are considered as ethically sensitive data. How these data are considered in the privacy scale within the same domain and between the different EU countries?
- To understand the complexity of machine learning model using Abstraction Hierarchy, a commonly used framework in cognitive engineering to visually represent the effect of ethically sensitive information on (i) Information Flow and (ii) Information Accuracy
- To understand the effect of removing of ethically sensitive information from criminal's record during the analysis process on the ML model accuracy.

To use visualisation techniques to provide awareness and insight for the analyst in the decision-making process.

Five research questions were posed to meet the five research objectives. The research questions within the context of the five objectives outlined the basis for the chapter's conclusions. The research questions stated in the following sections are addressed in each context exploring implication and application of the study.

7.3.1 Comparative analysis of "the interpretation of the principle of purpose limitation" across the EU: Belgium, Germany and the UK

The purpose of the research question 1 was to determine the importance of privacy law in the algorithmic decision-making process. To get a better picture of how analysts from different countries within the same domain understood the privacy law mainly focusing on "purpose limitation". Data is important for algorithmic decision-making as law enforcement agencies collect a large amount of data during the investigation process. The purpose of processing the data has to be specified at the point of data collection. Depending on how broad the specification of a purpose can be, it might give the controller flexibility to decide how the data should be processed. In criminal intelligence, data collected for one purpose might be useful or beneficial for another crime. How does the principle of purpose limitation work in the criminal domain? Do the different countries within the EU interpret the term "purpose limitation" in the same way? From our study, it was found that countries have different interpretations of the same term. The UK was more lenient compared to the other two countries. In the UK, data collected during the investigation of a specific offense can be shared with other police forces for investigations into other offenses, since this sharing is not

‘incompatible’ (contradictory) to the purpose for which the data was collected. In contrast to this, in Germany and Belgium, data collected during the investigation of a specific offense cannot be shared with other forces. Data collected should only be used for the specific offence it was collected for.

7.3.2 Development of a prototype privacy scale

The purpose of the research question 2 was to understand the information that is considered as ethically sensitive information. During the investigation process, a large amount of data will be collected by law enforcement agencies, some these information will be ethically sensitive information during the analysis process, so the selection of the data is very important (Guyon & Elisseeff, 2003; Shardlow, n.d.). According to privacy law, analysts are not allowed to use ethically sensitive information for analysis purposes. Therefore, they are removed or avoided during analysis. It is important to understand what information is considered as ethically sensitive information. However, what information is considered ethically sensitive information is not clear as it is contextually based. It depends on the individual, how they understand each piece of information. Additionally, ethically sensitive information can vary in its sensitivity. The choices of ethically sensitive information used can have an impact from two different perspectives: (i) information that leads to identifying an individual (personally identifiable information) and (ii) discriminatory (prejudice) towards certain minorities. In the police analysis, analysts will come across different prejudice and personally identifiable information, so it is very difficult to identify what attributes are more prejudice or personally identifiable information as the Data Protection Act has not given a precise list. During the analysis process, analysts have different mind sets and they will not consider to look at different types of attributes and weigh them accordingly so as not to make any harmful or discriminatory decision.

A proposed a privacy scale to advise analyst working in criminal intelligence. The designed and developed privacy matrix will give analysts a guideline to which attribute is prejudice and personally identifiable information based on their scale. An analyst can decide, based on their organization's legal position and code of conduct, whether or not to use the feature for analysis purposes. This privacy scale will help the analyst to be aware of the information based on their sensitivity level. From our study, it was found that even people within the same organization and the same workplace have a different viewpoints on the sensitivity of

data. An analyst can make their selection by looking at the matrix and the overall accuracy of the classification results.

The focus of the privacy scale was mainly on police analysis; however, this matrix can be used in other domains, not just on criminology domain. The attributes used in different domains such as finance or education will consist of similar personal information. The score that was proposed for single attributes can be used in other domains.

7.3.3 Modelling of ethically sensitive information using the abstraction hierarchy

The purpose of the research question 3 was to understand the complexity of the computational process. This chapter involved a visualisation of the model to show the complex processes in machine learning. Using the VALCRI system as a case study, I understood first what I need to represent to make the black box visible for inspection. To understand the complexity of the model, we used the abstraction hierarchy framework to analyse a complex socio-technical system, which is commonly used in the field of cognitive engineering. This could help the analyst understand the potential bias and discrimination in the algorithmic process.

Abstraction hierarchy is a diagram constructed through means-ends relations. A means-ends relation reveals the resources or constraints at one level that must be used for the satisfaction of resources or constraints at the next level up. Means-ends relation shows how-why relation to each other. The use of means-ends relation allows making structural relationships according to different levels of constraints. When looking for a reason for why one decision was made over others, we tend to consider the holistic properties of a system at a higher level of the abstraction. However, the reason for a certain decision could be because of the different processes within the system's component. As many components influence a certain outcome, it is difficult to explain a particular property for an outcome.

From our AH model, a functional relationship that represents key performance relationships was identified for grouping crime report according to their similarities. Based on the AH diagram, it was found that the similarity of the report is measured based on how much alike two reports are. Similarity, a measure in machine learning is a distance with dimensions representing features of the object. If the distance is small, the two reports have a high degree of similarity, whereas if the distance is large, the two report do not have many similarities. As a result, the feature plays an important role in identifying the similarity of reports.

Furthermore, from the AH diagram, we can see adhering with ethical values and legal principles is one of the aims of the VALCRI system. Semantic mapping principle (Bennett & Flach, 2011) was applied to make analysts aware of the ethically sensitive information.

7.3.4 Discrimination-aware machine learning

The purpose of the research question 4 was to determine the impact of using ethically sensitive information on the ML model accuracy. Based on the findings from chapter 4 on personal information, I experimented with the impact of including ethically sensitive information during the analysis process. Open-source data from ProPublica's COMPAS dataset was used for analysis purposes. The dataset consists of several attributes that are considered as personal information. The main purpose of this study was to understand the impact on accuracy when using ethically sensitive information during the investigation. From analysis, it was found that by including ethically sensitive information during the analysis process, I was able to acquire more accurate results. Earlier research by Angwin et al., (2016) showed removing the ethically sensitive information during analysis process does not always give fair and biased free results and the accuracy of the outcome was compromised. In Angwin et al. (2016) results were biased to certain ethnicity. Black defendants were incorrectly predicted to reoffend at a rate of 44.9%, nearly twice as high as their white counterparts at 23.5%; and white defendants were incorrectly predicted to not reoffend at a rate of 47.7%, nearly twice as high as their black counterparts at 28.0%.

In addition to accuracy, lack of awareness of ethically sensitive information being used was another problem that was currently faced by analyst. Algorithmic opacity problem hides the complexity of the process and people are not aware of the latent variable that results in a discriminatory outcome. Being aware that protected attributes are being used during the analysis process is important. Discrimination- awareness approach makes analysts aware of the use of ethically sensitive information during the analysis period. As a result, the analyst can decide to use or not use the information depending on their organisation privacy rules. Simple visualisation techniques were used to get insights and richer understanding that enables the analyst to be ethically aware and make an ethical decision. The visualization process was considered as a medium to provide transparency to the analysts during the analysis process. different simple visualization techniques such as bar charts, graphs and

interactive confusion matrix was used to support and provide guidance for the decision-making process.

7.4 Significance of the study

The study has shown the importance of ethics and privacy in the algorithms that are used for the decision-making process. The significance of this thesis is primarily in the field of Ethics, especially in Data Ethics field and to some extent in the Machine Learning field. The implication of this study can be divided into two parts:
(i) theoretical contributions and (ii) practical contributions.

In terms of theoretical contributions, based on the previous adoption, the studies have focused mainly on legal or privacy prospects or ethical perspective. However, when it comes to the machine learning approach, data ethics has a big impact on the issues discussed in this thesis. Data ethics helps to understand the potential ethical issues associated with collection and use of data specially ethically sensitive data. There is not much research regarding personal data. People are not aware of what is considered personal data because they are contextually based. In this study, instead of specifying that personal data need to be avoided, it discusses how this information can help to get a more accurate result in criminal intelligence analysis. different attributes that are considered as personal data to make analysts aware of these data were listed.

Besides that, in this thesis, I offered a narrative literature review that covers various aspects of machine learning, data ethics and privacy. The focus of the literature review was on understanding the issues that arose when developing the VALCRI system.

Regarding practical contributions, this study proposed a privacy scale, comparative analysis and visualization of the computational process. Based on the privacy scale and comparative analysis, the analyst should be able to be aware of ethically sensitive information. The visualization process will enhance awareness and help to get more insight into understanding ethical decision making.

7.5 Challenges

The concept of algorithmic ethics is new to the research. Until recent years, machine learning and ethics were seen as two different areas and not much was done collaboratively. When the VALCRI project started there was not much discussion or resources about ethics in the field of machine learning.

Initially, when the project started, ethicists and programmers had different viewpoints about algorithmic ethics. With more discussions and meetings, ethicists and programmers have started to talk with similar language and perspectives.

There were some conflicts between ethical principles. Compromise between some trade-off such as accuracy, fairness and privacy principles had to be made. In machine learning, it is difficult to incorporate all the ethical principles as principles contradict with each other. In our research, this conflict occurred between fair, accurate results vs privacy. In sensitive domains such as criminal intelligence, accuracy and fairness is as important as privacy.

During the data analysis process, analysts have to deal with a challenge regarding data and choices they made while selecting features, algorithms and transformation.
Generally, people working in the field of algorithmic ethics faced and still are facing a variety of challenges such as whether it is even possible to embed ethics; how to resolve conflicts between ethical framework, cultural challenges, algorithmic opacity, algorithmic bias and ethical rules.

7.6 Future directions

This thesis does not solve all of the problems nor fill all of the gaps regarding ethics in machine learning and data. Ample amount of research work is needed in this relatively new area, especially implementation work on the interaction between machine and human decision-making. As a result, there are several avenues for furthering this research work. The following questions arising from this thesis could be the subject of further research:

1. There is a difference in social values and ethical rules across cultures. As the focus of this thesis was on three countries within the same domain. A possible direction would be to study privacy principles in other countries within the EU and around the globe in the setting of a different domain such as medical and education.

2. The privacy scale proposed in this paper was evaluated by VALCRI end users. Further validation needs to be undertaken by other analysts and practitioners in the field to evaluate its usability and usefulness in the data analysis process in decision-making.

3. A visualization approach was taken to provide analysts some form of transparency. Visualization provides some awareness and insight into the outcome. Transparency is not just one thing that must be taken as a whole or not. Different analysts will have different levels of understanding of the outcome received by machine learning. In this thesis, simple visualization techniques were used to advise an analyst how visualization can enhance transparency. As future work, advanced visualization techniques can be done depending on individual analysts' needs, and analysts can decide what level of explanation they need.

References

Abate, T., & Krakovsky, M. (2018). Which is more fair: a human or a machine? Retrieved from https://engineering.stanford.edu/magazine/article/which-more-fair-human-or-machine

Al-Rubaie, M., & Chang, J. M. (2018). *Privacy Preserving Machine Learning: Threats and Solutions*. Retrieved from https://arxiv.org/pdf/1804.11238.pdf

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems*, *21*(4), 12–17. https://doi.org/10.1109/MIS.2006.83

Anderson, M., & Anderson, S. L. (2007). Machine Ethics : Creating an Ethical Intelligent Agent, *28*(4), 15–26.

Anderson, M., Anderson, S. L., & Armen, C. (2006). An approach to computing ethics. *IEEE Intelligent Systems*, *21*(4), 56–63. https://doi.org/10.1109/MIS.2006.64

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias. *ProPublica*, 1–16. https://doi.org/http://dx.doi.org/10.1108/17506200710779521

Ann Cavoukian. (2013). Privacy by design: The 7 foundational principles. Information and Privacy Commissioner of Ontario.

Article 29 Working Party. (2013). Opinion 03 / 2013 on purpose limitation. *Weisungen, Guidelines Etc*, (April), 1–70. Retrieved from http://ec.europa.eu/justice/data-protection/index_en.htm

Association of Chief Police Officers (ACPO). (2006). *Data Protection- Manual of Guidance Part 1: Standards*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/100376/13869_HO_guide_info_crime_victim.pdf

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*, 671–732. https://doi.org/10.15779/Z38BG31

Bennett, K. ., & Flach, J. . (2011). *Display and Interface Design: Subtle Science, Exact Art*. Boca Raton: CRC Press, Taylor and Francis Group.

Berendt, B., & Preibusch, S. (2014). Better decision support through exploratory discrimination-aware data mining: Foundations and empirical evidence. *Artificial Intelligence and Law*, *22*(2), 175–209. https://doi.org/10.1007/s10506-013-9152-0

Bundesministerium der Justiz und für Verbraucherschutz. (2014). Federal Data Protection Act. Retrieved from https://www.gesetze-im-internet.de/englisch_bdsg/englisch_bdsg.html

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Cadwalladr, C., & Graham-Harrison, E. (2018). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. Retrieved from

https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Calders, T., & Žliobaitė, I. (2013). Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures. In *Discrimination and Privacy in the Information Society Data Mining and Profiling in Large Databases*.

Carol Ormand. (2011). Developing Student Understanding of Complex Systems in the Geosciences. Retrieved from https://serc.carleton.edu/NAGTWorkshops/complexsystems/introduction.html

Centre for Internet and Human Rights. (2015). The Ethics of Algorithms: from radical content to self-driving cars - Final Draft Background Paper. *GCCS*, (1), 1–18. Retrieved from https://www.gccs2015.com/sites/default/files/documents/Ethics_Algorithms-final doc.pdf

Chainey, S., Tompson, L., & Uhlig, S. (2008). The Utility of Hotspot Mapping for Predicting Spatial Patterns of Crime. *Security Journal*, *21*(1–2), 4–28. https://doi.org/10.1057/palgrave.sj.8350066

Chalfin, A., Danieli, O., Hillis, A., Jelveh, Z., Luca, M., Ludwig, J., & Mullainathan, S. (2016). Productivity and Selection of Human Capital with Machine Learning. *American Economic Review*, *106*(5), 124–127. https://doi.org/10.1257/aer.p20161029

Cheng, K. M. (2018). *Rollins College Rollins Scholarship Online Predictive Analytics in the Criminal Justice System: Media Depictions and Framing Predictive Analytics in the Criminal Justice System: Media Depictions and Framing*. Retrieved from https://scholarship.rollins.edu/honors/62

Chris Albon. (2017). Feature Selection Using Random Forest. Retrieved from https://chrisalbon.com/machine_learning/trees_and_forests/feature_selection_using_random_forest/

Citron, D. K., & Pasquale, F. (2014). The Scored Society: Due Process for Automated Predictions. *Washington Law Review*, *89*, 101–133.

Cofone, I. N. (2019). *Algorithmic Discrimination Is an Information Problem*. *Hastings Law Journal* (Vol. 70). Retrieved from https://repository.uchastings.edu/hastings_law_journal/vol70/iss6/1

Creswell, J. W. (2013). *Qualitative inquiry and research design : choosing among five approaches*.

Cumbley, R., & Peter, C. (2008). What Is Personal Data? Retrieved from http://www.linklaters.com/Insights/Publication1403Newsletter/PublicationIssue20081001/Pages/PublicationIssueItem3513.aspx

Cutler, A., Pribić, M., & Humphrey, L. (2018). Everyday Ethics for Artificial Intelligence. *Ibm*, 33. Retrieved from www.ibm.com/legal/us/en/copytrade.shtml

Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*, *29*(3), 245–268. https://doi.org/10.1007/s13347-015-0211-1

Daniel Solove. (2014). What Is Sensitive Data? Different Definitions in Privacy Law. Retrieved from https://www.teachprivacy.com/sensitive-data-different-definitions-privacy-law/

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Retrieved April 17, 2019, from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Data Protection Act. (1998). Data Protection Act 1998. Statute Law Database.

Data Protection Commissioner. EU Directive 95/46/EC: The data protection directive (1995). Retrieved from https://www.dataprotection.ie/docs/EU-Directive-95-46-EC-Chapter-1/92.htm

Diakopoulos, N. (2014). Digital Journalism Algorithmic Accountability. https://doi.org/10.1080/21670811.2014.976411

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*

Directives 95/46/EC. (n.d.). Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. *Official Journal L 281 , 23/11/1995 P. 0031 - 0050;* Retrieved from https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:en:HTML

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. Retrieved from https://arxiv.org/pdf/1702.08608.pdf

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). *Fairness Through Awareness*. Retrieved from http://www.cs.toronto.edu/~zemel/documents/fairAwareItcs2012.pdf

Endert, A., Ribarsky, W., Turkay, C., Wong, B. L. W., Nabney, I., Blanco, I. D., & Rossi, F. (2017). The State of the Art in Integrating Machine Learning into Visual Analytics. *Computer Graphics Forum*, *36*(8), 458–486. https://doi.org/10.1111/cgf.13092

Endert, Alex, Fiaux, P., & North, C. (2012). *Semantic Interaction for Visual Text Analytics*. Retrieved from https://dl.acm.org/doi/10.1145/2207676.2207741

European Commission. (2018). Guidelines on transparency under Regulation 2016/79. Retrieved from https://ec.europa.eu/newsroom/article29/news.cfm?item_type=1360

European Commission Justice. (2010). Protection of personal data in Europe Union. Retrieved from http://ec.europa.eu/justice/data-protection/files/eujls08b-1002_-_protection_of_personnal_data_a4_en.pdf

European union agency for fundamental rights. (n.d.). Data protection | European Union Agency for Fundamental Rights. Retrieved from http://fra.europa.eu/en/data-protection

Federal Republic of Germany. (2009). Federal Data Protection Act (BDSG). Retrieved October 11, 2019, from https://www.gesetze-im-internet.de/englisch_bdsg/englisch_bdsg.html

Federal Trade Commission. (2013).The Privacy Challenges of Big Dat: A View from the Lifeguard's Chair. Retrieved from http://www.ftc.gov/os/caselist/1023136/111024googlebuzzcmpt.pdf

Floridi, L., & Taddeo, M. (2016). What is data ethics? https://doi.org/10.1098/rsta.2016.0360

Gillespie, T. (2012). The relevance of algorithms. *Media Technologies: Essays on Communication, Materiality, and Society*, (Light 1999), 167–194. https://doi.org/10.7551/mitpress/9780262525374.003.0009

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a right to explanation. Retrieved from https://arxiv.org/pdf/1606.08813v3.pdf

Gov.uk. (2019). *Stop and search*. Retrieved from https://www.ethnicity-facts-figures.service.gov.uk/crime-justice-and-the-law/policing/stop-and-search/latest

Guthrie Ferguson, A. (2012). Predictive Policing and Reasonable Suspicion. *Emory Law Journal*. Retrieved from http://ssrn.com/abstract=2050001

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research (JMLR)*, *3*(3), 1157–1182. https://doi.org/10.1016/j.aca.2011.07.027

Hall, P., & Gill, N. (2018). Debugging the Black-Box COMPAS Risk Assessment Instrument to Diagnose and Remediate Bias.

Hannah Couchman. (2019). Policing by Machine. Retrieved from https://www.libertyhumanrights.org.uk/sites/default/files/LIB 11 Predictive Policing Report WEB.pdf

Harari, Y. N. (2018). *21 lessons for the 21st century*. Retrieved from https://books.google.co.uk/books/about/21_Lessons_for_the_21st_Century.html?id=ar44DwAAQBAJ&printsec=frontcover&source=kp_read_button&redir_esc=y#v=onepage&q&f=false

Hasselbalch, G., & Tranberg, P. (2017). *Data Ethics — The New Competitive Advantage* (Vol. 91). Retrieved from https://dataethics.eu/wp-content/uploads/DataEthics-UK-original.pdf

Hearst, M. (2009). *Search user interfaces*. Cambridge University Press. Retrieved from http://searchuserinterfaces.com/book/

Hepenstal, S., Kodagoda, N., Zhang, L., Paudyal, P., & Wong, B. L. W. (2019). *Algorithmic Transparency of Conversational Agents* (Vol. 11). Retrieved from http://ceur-ws.org/Vol-2327/IUI19WS-IUIATEC-1.pdf

Hill, K. (2012). How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did. Retrieved from https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured- out-a-teen-girl-was-pregnant-before-her-father-did/

House of Common. (2018a). *Algorithms in decision- making*. Retrieved from www.parliament.uk/science

House of Common. (2018b). Algorithms in decision-making. Retrieved from https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/351/35105.htm

Huber,Rose . (2014). “Mosaic Effect” Paints Vivid Pictures of Tech Users’ Lives, Felten Tells Privacy Board | Woodrow Wilson School of Public and International Affairs. Retrieved from http://wws.princeton.edu/news-and-events/news/item/mosaic-effect-paints-vivid-pictures-tech-users-lives-felten-tells-privacy

Hullman, J. (2019). The purpose of visualisation is insight, not picture: An interview with Ben Shneiderman. Retrieved from http://interactions.acm.org/blog/view/the-purpose-of-visualization-is-insight-not-pictures-an-interview-with-ben

Hurwitz, J., & Kirsch, D. (2018). *Machine Learning*. Retrieved from http://www.wiley.com/go/permissions.

IEB. (2017). *"Independent Ethics Board Second Report on VALCRI Project", the Midterm Report to European Commission January*.

Information Commissioner's Office. (2015). Key definitions of the Data Protection Act. *Internet*. Retrieved from https://ico.org.uk/for-organisations/guide-to-data-protection/key-definitions/

Information Commissioner's Office. (2017). General Data Protection Regulation (GDPR), 43 pages. Retrieved from https://ico.org.uk/media/for-organisations/data-protection-reform/overview-of-the-gdpr-1-13.pdf

International Labour Office. (1992). *Workers' privacy. Part I : protection of personal data.* International Labour Office. Retrieved from https://www.ilo.org/global/topics/safety-and-health-at-work/normative-instruments/code-of-practice/WCMS_107797/lang--en/index.htm

Jason, B. (2016). Overfitting and Underfitting With Machine Learning Algorithms. Retrieved from https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/

Jason Brownlee. (2017). Why One-Hot Encode Data in Machine Learning? Retrieved from https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/

Jeong, D. H., Ziemkiewicz, C., Fisher, B., Ribarsky, W., & Chang, R. (2009). iPCA: An Interactive System for PCA-based Visual Analytics. *Computer Graphics Forum*, *28*(3), 767–774. https://doi.org/10.1111/j.1467-8659.2009.01475.x

Jester, A. M. B.-, Casselman, B., & Goldstein, D. (2015). The new science of writing. In *The Reader's Brain* (pp. 10–28). https://doi.org/10.1017/cbo9781316178942.002

Julia Angwin, Jeff Larson, Surya Mattu, & Lauren Kirchner. (2016). Machine Bias. Retrieved April 17, 2019, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Kraemer, F., van Overveld, K., & Peterson, M. (2011). Is there an ethics of algorithms? *Ethics and Information Technology*, *13*(3), 251–260. https://doi.org/10.1007/s10676-010-9233-7

Larson, J. (2016). COMPAS dataset. Retrieved from https://github.com/propublica/compas-analysis

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016a). How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica.Org*, pp. 1–17. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Lee, D. (2019). AI helps clear cannabis conviction backlog. Retrieved from https://www.bbc.co.uk/news/av/technology-48067725/ai-helps-clear-cannabis-conviction-backlog

Lintern, G. (2013). Tutorial: Work Domain Analysis, (1999). https://doi.org/10.1201/b14774

Lipton, Z. C. (2016). The Mythos of Model Interpretability. Retrieved from https://arxiv.org/pdf/1606.03490.pdf

Liu, S., Dissanayake, S., Patel, S., Dang, X., Mlsna, T., Chen, Y., & Wilkins, D. (2014). Learning accurate and interpretable models based on regularized random forests regression. *BMC Systems Biology*, *8*(Suppl 3), S5. https://doi.org/10.1186/1752-0509-8-S3-S5

Marquenie, T., Coudert, F., Duquenoy, P., & Paudyal, P. (2017). Roadmap for the resolution of ethical and human rights issues in automated data analysis and extraction computations in VALCRI. VALCRI White Paper Series. Retrieved from https://lirias.kuleuven.be/1711906?limo=0

Marsland, S. (2009). *Machine learning : an algorithmic perspective*. CRC Press.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McCallister, E., Grance, T., & Kent, K. (2010). Guide to protecting the confidentiality of personally identifiable information (PII). *Special Publication 800-122 Guide*, 1–59. https://doi.org/10.6028/NIST.SP.800-122

McLeod, S. (2008). Prejudice and Discrimination in Psychology. Retrieved from https://www.simplypsychology.org/prejudice.html

Michael Josephson. (2016). Fairness - Exemplary Business Ethics & Leadership. Retrieved from http://josephsononbusinessethics.com/2010/12/fairness/

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society*, *2016*(December), 68. https://doi.org/10.1177/2053951716679679

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. https://doi.org/10.1145/3287560.3287574

Molnar, C. (2019). *Interpretable Machine Learning*. Reterived from https://christophm.github.io/interpretable-ml-book/

Montavon, G., Samek, W., & Müller, K.-R. (2017). Methods for Interpreting and Understanding Deep Neural Networks. https://doi.org/10.1016/j.dsp.2017.10.011

Moor.J.H. (1985). What is Computer Ethics? Reterived from https://www.jstor.org/stable/24436819

Moses, L., & Chan, J. (2014). Using big data for legal and law enforcement decisions: Testing the new tools. *University of New South Wales Law Journal, The*, *37*(2), 643.

Murphy, M. H. (2017). Algorithmic Surveillance: The Collection Conundrum. *Computers & Technology*.

Naikar, N. (2013). *Work domain analysis : concepts, guidelines, and cases*. CRC Press.

Narayanan, A., & Shmatikov, V. (2008). *Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)*. Retrieved from https://arxiv.org/pdf/cs/0610105.pdf

Narayanan, A., & Shmatikov, V. (2010). Myths and fallacies of "personally identifiable information." *Communications of the ACM*, *53*(6), 24. https://doi.org/10.1145/1743546.1743558

Nass, S. J., Levit, L. A., Gostin, L. O., & Rule, I. (2009). The Value and Importance of Health Information Privacy. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK9579/

Nath, S. (2006). Crime pattern detection using data mining. *Web Intelligence and Intelligent Agent Technology*(954), 4. https://doi.org/10.1109/WI-IATW.2006.55

Neal E. Boudett. (2016). Tesla Faults Brakes, but Not Autopilot, in Fatal Crash. Retrieved from https://www.nytimes.com/2016/07/30/business/tesla-faults-teslas-brakes-but-not-autopilot-in-fatal-crash.html

Nguyen, P. H., Xu, K., Wheat, A., Wong, B. L. W., Attfield, S., & Fields, B. (2016). SensePath: Understanding the Sensemaking Process Through Analytic Provenance. *IEEE Transactions on Visualization and Computer Graphics*, *22*(1), 41–50. https://doi.org/10.1109/TVCG.2015.2467611

Nissenbaum, H. (2004). PRivacy as Contextual Integrity. Retrieved from https://crypto.stanford.edu/portia/papers/RevnissenbaumDTP31.pdf

Norman, D. (1986). *The Design of Everyday Things*. https://doi.org/10.15358/9783800648108

O'Neil, C. (2016). *Weapons of math destruction : how big data increases inequality and threatens democracy*. Penguin UK

Orum, A. M. (1992). A case for the Case Study. *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*, (September), 202–207. https://doi.org/10.1016/B978-0-08-097086-8.44002-X

Pasquale, F. (2015). The Black Box Society. *Cambridge, MA: Harvard University Press*, *36*, 32. Retrieved from http://www.hup.harvard.edu/catalog.php?isbn=9780674368279

Paudyal, P., Rooney, C., Kodagoda, N., Wong, B. L. W., Duquenoy, P., & Qazi, N. (2017). How the Use of Ethically Sensitive Information Helps to Identify Co-Offenders via a Proposed Privacy Scale: A Pilot Study. In *2017 European Intelligence and Security Informatics Conference (EISIC)* (pp. 164–164). IEEE. https://doi.org/10.1109/EISIC.2017.35

Paudyal, P., & William Wong, B. L. (2018). Algorithmic Opacity: Making Algorithmic Processes Transparent through Abstraction Hierarchy. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 192–196. https://doi.org/10.1177/1541931218621046

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08* (p. 560). https://doi.org/10.1145/1401890.1401959

Pichai, S. (2018). AI at Google: our principles. Retrieved from https://ai.google/responsibilities/responsible-ai-practices/

Pohl, M., Winter, L.-C., Pallaris, C., Attfield, S., & Wong, B. L. W. (2014). Sensemaking and Cognitive Bias Mitigation in Visual Analytics. In *2014 IEEE Joint Intelligence and Security Informatics Conference* (pp. 323–323). IEEE. https://doi.org/10.1109/JISIC.2014.68

Privacy Commission. (2009). *Royal Decree Implementing the Law of 8 December 1992 on the protection of privacy in relation to the processing of personal data - consolidated version of September 2008*. Retrieved from https://www.privacycommission.be/sites/privacycommission/files/documents/Royal_Decree_20

01.pdf

Raab, C., & Goold, B. (2011). *Protecting information privacy*. Retrieved from www.equalityhumanrights.com

Rahman, M. M., & Davis, D. N. (2013). Addressing the Class Imbalance Problem in Medical Datasets. *International Journal of Machine Learning and Computing*, *3*, 224–228. https://doi.org/10.7763/IJMLC.2013.V3.307

Rasmussen, J. (1985). The role of hierarchical knowledge representation in decisionmaking and system management. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-15*(2), 234–243. https://doi.org/10.1109/TSMC.1985.6313353

Reising, D. (2000). The Abstraction Hierarchy and its Extension beyond Process Control. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*(1), 194–197. https://doi.org/10.1177/154193120004400152

Reising, D. V. C., & Sanderson, P. M. (2002). Ecological Interface Design for Pasteurizer II: A Process Description of Semantic Mapping Dal.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier Marco. https://doi.org/10.1145/2939672.2939778

Richterich, A. (2018). *The Big Data Agenda: Data Ethics and Critical Data Studies*. *The Big Data Agenda: Data Ethics and Critical Data Studies*. https://doi.org/10.16997/book14

Rieke, A., Bogen, M., & Robinson, D. G. (2018). Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods. Retrieved from https://www.omidyar.com/sites/default/files/file_archive/Public Scrutiny of Automated Decisions.pdf

Rigano, C. (2018). *Using Artificial Intelligence to Address Criminal Justice Needs*. Cham. Retrieved from https://www.nij.gov/journals/280/Pages/using-artificial-intelligence-to-address-criminal-justice-needs.aspx

Rippert, S., Weimer, K. A., & Llp, R. S. (2009). Data Protection: Germany. Retrieved from http://whichlawyer.practicallaw.com/9-385-8462?qp=&qo=&q=

Royal Society, T. (n.d.). Machine Learning: The Power and Promise of Computers that learns by Example. Retrieved from https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf

Sacha, D., Jentner, W., Zhang, L., Stoffel, F., & Ellis, G. (2017). *Visual Comparative Case Analytics*. Retrieved from http://eprints.mdx.ac.uk/21853/1/visual-crime-case.pdf

Sacha, D., Jentner, W., Zhang, L., Stoffel, F., Ellis, G., & Keim, D. (2017). *Applying Visual Interactive Dimensionality Reduction to Criminal Intelligence Analysis*. Retrieved from https://scibib.dbvis.de/uploadedFiles/VALCRIWP2017011InteractiveVisualDimensionReduction.pdf

Sacha, D., Sedlmair, M., Zhang, L., Lee, J. A., Peltonen, J., Weiskopf, D., … Keim, D. A. (2017). What you see is what you can change: Human-centered machine learning by interactive visualization. *Neurocomputing*, *268*, 164–175. https://doi.org/10.1016/J.NEUCOM.2017.01.105

Sacha, D., Zhang, L., Sedlmair, M., Lee, J. A., Peltonen, J., Weiskopf, D., … Keim, D. A. (2017).

Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 241–250. https://doi.org/10.1109/TVCG.2016.2598495

Sánchez-Monedero, J., & Dencik Sanchez-Monederoj, L. (2018). *How to (partially) evaluate automated decision systems*. Retrieved from https://datajusticeproject.net/wp-content/uploads/sites/30/2018/12/WP-How-to-evaluate-automated-decision-systems.pdf

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2016). Automation, Algorithms, and Politics | When the Algorithm Itself is a Racist: Diagnosing Ethical Harm in the Basic Components of Software. *International Journal of Communication*, *10*(0), 19.

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. *Handbook of Human Factors and Ergonomics*, *2*, 1926–1943. https://doi.org/10.1207/s15327108ijap0204_5

Saul McLeod. (2019). Case Study Method. Retrieved from https://www.simplypsychology.org/case-study.html

Schlabach, G. R. (2014). Privacy in the Cloud: The Mosaic Theory and the Stored Communications Act. Retrieved from http://www.theatlantic

Schwenke, C., & Schering, A. (2007). True Positives, True Negatives, False Positives, False Negatives. In *Wiley Encyclopedia of Clinical Trials*. Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9780471462422.eoct021

Shardlow, M. (n.d.). An Analysis of Feature Selection Techniques. *Studentnet.Cs.Manchester.Ac.Uk*, 1–7. Retrieved from https://studentnet.cs.manchester.ac.uk/pgt/COMP61011/goodProjects/Shardlow.pdf

Silver, D., Schrittwieser, J., Simonyan, K., Nature, I. A.-, & 2017, U. (2009). Mastering the game of Go without human knowledge. *Nature.Com*. Retrieved from https://www.nature.com/articles/nature24270?sf123103138=1

Smith, B., & Shum, H. (2018). *The Future Computed Artificial Intelligence and its role in society*. Retrieved from https://blogs.microsoft.com/uploads/2018/02/The-Future-Computed_2.8.18.pdf

Stahl, B. C., Timmermans, J., & Mittelstadt, B. D. (2016). The Ethics of Computing: A Survey of the Computing-Oriented Literature. https://doi.org/10.1145/2871196

Starr, D. (2018). Current use cases for machine learning in healthcare. Retrieved July 22, 2019, from https://azure.microsoft.com/en-gb/blog/current-use-cases-for-machine-learning-in-healthcare/

Stefaan Verhulst. (2015). Big and Open Linked Data (BOLD) in government: A challenge to transparency and privacy? Retrieved from http://thegovlab.org/big-and-open-linked-data-bold-in-government-a-challenge-to-transparency-and-privacy/

Sutton, R. S., & Barto, A. G. (2017). *Reinforcement Learning: An Introduction*. Retrieved from http://incompleteideas.net/book/bookdraft2017nov5.pdf

Sweeney, L. (2013). Discrimination in online ad delivery. *Communications of the ACM*, *56*(5), 44–54. https://doi.org/10.1145/2460276.2460278

Tang, Jian, Liu, J., Zhang, M., & Mei, Q. (2016). Visualizing Large-scale and High-dimensional Data.

https://doi.org/10.1145/2872427.2883041

Tang, Jiliang, Alelyani, S., & Liu, H. (2014). Feature Selection for Classification: A Review. *Data Classification: Algorithms and Applications*, 37–64. https://doi.org/10.1.1.409.5195

Tashakkori, A., & Teddlie, C. (2003). *Handbook of Mixed Methods in Social & Behavioral Research.* Thousand Oaks: Sage.

The Economist. (2016). From not working to neural networking. Retrieved November 13, 2019, from https://www.economist.com/special-report/2016/06/23/from-not-working-to-neural-networking

The European Commission. (n.d.). *What is personal data?* Reterived from https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en

The Royal Society. (2015). *Protecting privacy in practice The current use, development and limits of Privacy Enhancing Technologies in data analysis Protecting privacy in practice: The current use, development and limits of Privacy Enhancing Technologies in data analysis*. Retrieved from https://royalsociety.org/-/media/policy/projects/privacy-enhancing-technologies/privacy-enhancing-technologies-report.pdf

The White House. (2016). *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights Executive Office of the President Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*. Retrieved from https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf

Timothy Revell. (2017). AI detective analyses police data to learn how to crack cases. Retrieved from https://institutions.newscientist.com/article/mg23431254-000-ai-detective-analyses-police-data-to-learn-how-to-crack-cases/

Treacy, B., Osmani, F., Smith, W., & Firth, R. (2016). Preparing for Change: Europe's Data Protection Reforms Now a Reality. Retrieved from https://www.huntonak.com/images/content/3/4/v3/3461/Germany-The-International-Comparative-Legal-Guide-to-Data-Protec.pdf

Turkay, C., Laramee, R., & Holzinger, A. (2017). On the Challenges and Opportunities in Visualization for Machine Learning and Knowledge Extraction: A Research Agenda. *LNCS*, *10410*, 191–198. https://doi.org/10.1007/978-3-319-66808-6

Tutt, A. (2016). *An FDA for Algorithms*. *SSRN*. https://doi.org/10.2139/ssrn.2747994

USACM. (2017). Principles for Algorithmic Transparency and Accountability, 1–2. Retrieved from https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

VALCRI. (2014). VALCRI - Visual Analytics for Sense-making in CRiminal Intelligence analysis. Retrieved from http://valcri.org/

Valentino-DeVries, J., Singer, N., Keller, M. H., & Krolik, A. (2018). Your Apps Know Where You Were Last Night, and They're Not Keepting it Secret (5). *The New York Times*. Retrieved from https://www.nytimes.com/interactive/2018/12/10/business/location-data-privacy-apps.html

van den Hoven, J. (2008). Information technology, privacy, and the protection of personal data. *Information Technology and Moral Philosophy*. https://doi.org/10.1017/CBO9780511498725.016

van Otterlo, M. (2013). A machine learning view on profiling. In *Privacy Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology* (pp. 41–64). https://doi.org/10.4324/9780203427644

Veale, M. (2019). Algorithms in the Criminal Justice System. Retrieved from https://epic.org/algorithmic-transparency/crim-justice/

Wagner, B. (2016). Algorithmic regulation and the global default : Shifting norms in Internet technology. *Etikk i Praksis - Nordic Journal of Applied Ethics*, *10*(1), 1–9. https://doi.org/10.5324/eip.v10i1.1961

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M., & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*, *12*(4), e0174944. https://doi.org/10.1371/journal.pone.0174944

Westermann, H. (2018). Change of Purpose - The effects of the Purpose Limitation Principle in the General Data Protection Regulation on Big Data Profiling. Retrieved from http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8941820&fileOId=8943991

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). *The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions*. Retrieved from www.aaai.org

Williams, Brooks, & Shmargad. (2018). How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications. *Journal of Information Policy*, *8*, 78. https://doi.org/10.5325/jinfopoli.8.2018.0078

Wodecki, A., Allen, G. C., Horowitz, M. C., Kania, E. B., Scharre, P., Wilson, H. J., … Borana, J. (2017). Explainable Artificial Intelligence (XAI) The Need for Explainable AI. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *58*(2), 4. https://doi.org/10.1111/fct.12208

Wood, M., Hales, J., Purdon, S., Sejersen, T., & Hayllar, O. (2009). *A test for racial discrimination in recruitment practice in British cities*. Retrieved from www.ecu.ac.uk/wp-content/uploads/2014/07/unconscious-bias-and-higher-education.pdf

Worsley, M. D. (1987). The UK data protection act 1984 and international writers. *IEEE Transactions on Professional Communication*, *PC30*(3), 208. https://doi.org/10.1109/TPC.1987.6449077

Woulds, J. (2004). A Practical Guide to the Data Protection Act. Reterived from https://www.ucl.ac.uk/constitution-unit/sites/constitution-unit/files/118.pdf

Zarsky, T. (2013). Transparent Predictions. *University of Illinois Law Review*, (4), 1503–1569

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

Zheng, A. (2015). Evaluating Machine Learning Models. Retrieved from https://www.oreilly.com/ideas/evaluating-machine-learning-models/page/3/evaluation-metrics

Ziewitz, M. (2015). Governing Algorithms Myth, Mess, and Methods. *Science, Technology, &*

Human Values.

Zliobaite, I. (2015). A survey on measuring indirect discrimination in machine learning. *ACM Journal Name*, *0*(0). Retrieved from https://arxiv.org/pdf/1511.00148.pdf