Responsible AI

The Praxis of AI and Data Protection Management: Negotiating Innovation and FAT Principles

Maria Chiara Addis

Salford Business School

University of Salford

PhD Student ID@00462630

Supervisors: Dr Maria Kutar & Dr Gordon Fletcher

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

August 2021

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GLOSSARY

List of terms and definitions used in this thesis

Artificial General Intelligence (AGI)

Artificial General Intelligence (AGI), or Superintelligence or Singularity, is a potentially superior version of AI, capable of self-evolving and of surpassing human intelligence.

Artificial Intelligence (AI)

AI is the capacity of a machine to perform or think like a human being. It is distinguished from Natural Intelligence (NI) of biological systems.

AI Ethics

The field that encompasses values and principles concerning the development and use of AI. The area is currently mainly focused on exploring the development of the technology, data and biases.

Augmented AI

Human-centred model. AI is generally used to support human decisions, not to replace them.

Autonomous AI

AI systems that make decisions without human interference.

Biometrics

Personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person (e.g., facial images or fingerprints).

CRAIDA Framework

Critical AI&DP framework. Theoretical framework developed to critically theorize the context of an organisation.

Critical Theory of Technology (CTT)

Critical theory of technology was created by Andrew Feenberg drawing upon philosophy of technology and constructivist technology studies. Influenced by the Frankfurt School, Heidegger, and social constructivism, his theory considers technologies and technological systems at different levels. Technology is considered socially constructed and instrumental to modern hegemonies.

Controller

The natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of processing personal data (GDPR).

Data Protection Act (DPA) 2018 and UK GDPR

Data Protection Act 2018 replaced the Data Protection Act 1998, defining the UK DP framework. It contains GDPR specifications and derogations, law enforcement processing, and a separate

	regime for national security and intelligence Services. The UK GDPR is the retained EU law version of the GDPR. The UK GDPR and an amended version of the DPA 2018 are now the main DP legislative texts in the UK.
Data Subject	An identifiable natural person who can be identified (directly or indirectly) by reference to a name, identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person (GDPR).
Digital Innovation	The application of digital technology to business problems.
Data Protection (DP)	The right to the protection of personal data, a human right in the European tradition. Convention 108 (European Council, 1981) was the first legislation on DP approved at European level.
Data Protection Impact Assessment (DPIA)	Tool used to evaluate the potential consequences of data processing in high-risk processing (i.e., process performed using new technologies).
Data Protection Officer (DPO)	DP independent expert who informs, advises and monitors GDPR compliance.
Fairness, accountability and transparency (FAT) principles	Research on AI and fairness, accountability and transparency has grown amongst AI/ML researchers in the last few years. FAT principles are also included in the GDPR.
Framework	A structure that provides the support for a system or a set of concepts.
GDPR	General Data Protection Regulation. EU legislative act regulating Data Protection. The Regulation became enforceable in May 2018.
Human-Computer/Machine Interaction	Research area that focuses on the interaction between human/user and computer/machine.
Human in the Loop	Human that interacts with machines in a ML model and makes decisions on top of predictions.

Know your Customer (KYC)	Guidelines on the identification, suitability and risk of customers.
Learning Analytic	Collection and analysis of data of learners aiming at improving their learning.
Machine Learning (ML) and Prediction	Subset of AI. ML learns from experience and examples, improves, and makes predictions. By using available data (usually from past and real-time events), ML algorithms predict/guess hidden or missing information.
Model	A practical tool that describes how different parts of an organisation can work together successfully
Narrow AI	Weak or Narrow AI is a type of AI that solves issues in specific domains using methods from other fields (i.e., statistics).
Natural Language Processing (NLP)	Subset of AI. It recognises patterns of words and sentences in a text, "understands" the content, and extracts information from it.
New Technologies	New kinds of technologies that alter the way something is produced or performed.
Open Banking	Open Banking is a banking practice that allows third-party entities open access to consumer banking data via application programming interfaces (APIs). Open Banking regulations promote the use of open-source technology, transparency, and wide interoperability between different subjects.
OSINT	Open-Source intelligence gathering and analysis accessible data in publicly available sources.
Privacy	Human right intended as the right to have private and family life, home, and communications respected.
Processor	A natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller.
Profiling	The use of personal data to analyse or predict aspects concerning a natural person (i.e., performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements).
RAIDIS Management Model	Responsible AI&DP Management Model. The model is a practical tool for organisations adopting responsible AI

management. The model includes different elements in the organisational context (technology, people, processes, stakeholders, decision-making) to be considered for an effective responsible management.

RAIDIS Maturity Model
Responsible AI&DP Maturity Model. The model illustrates how AI and DP can be included within an organisational strategy. It defines five stages of organisational maturity that indicate the evolution in the adoption and use of AI and DP.

Reinforcement Learning
Subset of ML based on the interaction of an agent with his environment. Learning is the result of external inputs, continuous interactions, decisions, rewards, learning and adaption. The goal of the agent is to learn the consequences of its decisions, such as which moves are important in winning a game, and to use this learning to find strategies that maximise its rewards.

Responsible AI Officer (RAO)
Role in charge of the full AI innovation cycle within an organisation.

Responsible Research and Innovation (RRI)
RRI is an EU governance framework for research and innovation. It is a key action of Horizon 2020, the EU financial instrument implementing research and innovation policy.

Semi-supervised Learning
Subset of ML. In semi-supervised learning only part of the data sets is labelled. Particularly useful in complex experiments, this method includes probabilistic models, graph-based semi-supervised learning, and transductive Support Vector Machines.

SHERPA Project
Shaping the Ethical Dimensions of Smart Information Systems is an EU-funded project which analyses how AI and big data analytics impact ethics and human rights.

Subject Matter Experts (SMEs)
Experts in a specific area or topic.

Student Progression Administrator (SPA)
Education staff supporting students in their learning.

Supervised Leaning
Subset of ML. Supervised Learning algorithms are trained using datasets labelled according to groups or categories. The system

learns what is included in that group, creates rules to understand the environment, adapt, and make predictions.

Unsupervised Learning

Unsupervised Learning algorithms do not learn via predefined categories but are set free to find similar characteristics in the dataset. The most common unsupervised learning task is clustering.

ACKNOWLEDGEMENTS

I would firstly like to thank the Salford Business School. The support given through the Pathway to Excellence Research scholarship enabled me to carry out this work. I will be forever grateful for this privilege.

I would also like to thank my supervisors Dr Maria Kutar and Dr Gordon Fletcher.

The completion of this PhD thesis would not have been possible in particular without the guidance, support, wise advice, expertise, encouragement, and deep care I received from Maria, my main PhD supervisor and former M.Sc. dissertation supervisor. She was there for me, always. Thank you so much.

I would like to thank the student support staff of the school and my fellow PhD colleagues at the university for their support and encouragement. To those who worked so hard for their doctorates, and to those who spent their days and some of their nights, at the Salford University Library. I wish you all the best. You deserve it.

I would also like to thank the participants of this study for taking the time to share their experiences, knowledge and interesting thoughts with me.

I am grateful to my friends for their continuous support, concern and care while completing this research. A special thanks to Sheila Ross and Sara Shepherd.

My profound gratitude goes to Christian Klesse. His love, deep care, patience, support, humour, and intelligence have been essential in this journey and in a difficult moment of my life. Grazie mille.

Finally, I extend my gratitude to my family for believing in me throughout my life. Their love and support have given me the strength and determination for facing many challenges. A special thanks to my sister Martina. Without her continuous love and help this work would not be here. This thesis is dedicated to her.

DECLARATION OF ORIGINALITY

This thesis is submitted under the University of Salford code of practice for the conduct of postgraduate research degree programmes. Some parts of this research have been published in refereed conference proceedings prior to the submission of this thesis.

I hereby declare that this thesis is my unaided work and is being submitted for the award of degree of Doctor of Philosophy at the University of Salford. There is no portion of thesis that had been submitted anywhere for the award of any academic qualification.

Addis

Maria Chiara Addis

29th August 2021

Supervisor: Maria Kutar

ABSTRACT

The increasing deployment of Artificial Intelligence applications has sparked a debate on its possible uses and potential problems, and many questions on the protection of personal data have emerged. The General Data Protection Regulation (GDPR) imposed new requirements for organisations handling personal data, and the implications for organisations managing AI technologies are particularly significant. Whereas much research focuses on algorithmic biases and the development of AI, this research explores other important concerns arising from the uses of personal data during the introduction of AI, which impact on individuals and organisations. It investigates innovation in different organisational contexts and how people perceive, understand and apply AI, data protection and FAT principles (fairness, accountability and transparency).

Drawing on responsible research and innovation (RRI) and Feenberg’s critical theory of technology, the research investigates the praxis of AI and GDPR management within UK organisations, examining the interplay between AI, data protection and FAT principles.

The methodology comprises a multi method approach, employing a survey of experts and dual case studies of organisations implementing responsible AI projects. This research investigates organisational practices and people's agency, providing in-depth analysis of values, power dynamics, experience, understanding, perceptions, and difficulties of various stakeholders (leaders, senior managers, data protection and ML experts) in their specific contexts, all of which shapes and constructs this ambivalent technology.

The research indicates that GDPR is often misinterpreted, there is limited understanding of AI and its specific risks, and there are diverse perceptions of the relevance of FAT principles. Discussion on ethics is usually focused on data and activities conducted prior to the implementation of new AI systems. Internal processes and personal data created by AI are generally unconcerned by discourse on responsible innovation. External partners raise special concerns around compliance and unethical practices.

This research critically reflects upon these flaws, identifies rarely discussed problems that obstruct responsible innovation and defines areas for innovation. Explaining how roles, positionality and personal experiences can impact management decisions regarding AI implementation, the research proposes an approach to AI innovation studies that foregrounds the active role of people in shaping technology. These insights are systematised in the creation of a critical AI and data protection management model aimed at supporting organisations to understand and address specific challenges, risks, and benefits in their responsible management. The research thereby offers leaders and senior managers important instruments for increasing awareness and control while using AI to process personal data. Highlighting the multilevel and multidisciplinary aspects of AI management, unveiling the complexities around ML predictions and decision making, and showing innovative potentials residing within the GDPR, this further contributes important insights to business and management studies and to interdisciplinary debates on AI, data protection, and organisational ethics.

CHAPTER 1: INTRODUCTION

1.1 Introduction

The research presented in this thesis focuses on Artificial Intelligence (AI), data protection (DP), and the concepts of fairness, accountability, and transparency (FAT).

Whereas much research explores the development of AI and biases in data, this research shifts the focus onto how the technology is used. By exploring the experience of the individuals and the praxis within the UK organisations that are introducing AI systems, this research investigates innovation in different contexts, identifies exact risks and potential within specific areas, and provides some models that can guide organisations in their responsible AI and DP management.

This first chapter begins with a focus on the context, which illustrates some of the concerns arising from the development and application of AI, the debate around fairness, accountability, and transparency of the algorithms, the major events occurring within DP legislation (e.g., GDPR), and some information on digital innovation and the challenges posed by AI. It then elucidates the motivation for conducting the research, its impact, and its contribution to the management of AI in practice, and following that, it presents the research question, aims and objectives. After presenting the theoretical framework and its origin, the chapter then illustrates the research methods and ends with some definitions of terms used throughout the research.

1.2 Context

The context of this research encompasses three different areas:

- Artificial Intelligence (AI) technology, potential and concerns.
- Data protection legislation, the GDPR (the European General Data Protection Regulation 2016/679), and the FAT (fairness, accountability, and transparency) Principles.
- The use of AI technologies in innovation management by UK organisations.

In the following section, the peculiarities and risks linked to AI, the GDPR and FAT principles, and their significance for innovation management will be presented.

1.2.1 AI

The rapid development of Artificial Intelligence (AI) technologies, Big Data and Data Economy has sparked the debate on the potential of, and concerns about, these technologies (Ananny & Crawford, 2018; Bird et al., 2016; Floridi et al., 2018; Iansiti & Lakhani, 2020a; Morley et al., 2020; Skirpan & Gorelick, 2017; Veale & Edwards, 2018; Zuboff, 2019). AI appears to have the potential to impact on the rights and freedoms of individuals, disrupt existing social, economic, and political orders, and enhance or limit economic possibilities. More than with other technologies, the progress of AI is creating a unique dichotomy, being seen as either capable of solving most of the problems of humanity, or of destroying it. While some argue that AI is not different from other technologies and should not be differently regulated, others see its inscrutability, autonomy, and unpredictability as unique, and request specific guidelines and regulations. New questions have emerged amongst academics, practitioners, and within the wider public (Borgesius, 2018; Cath et al., 2017; Crawford, 2017; Greenfield, 2018; O'Neil, 2016a; Sobel, 2020; TUC, 2021b, 2021a; Whittaker et al., 2018).

Algorithms are filtering knowledge, creating new taxonomies, selecting what to read, whom to date, who deserves a loan, and who can be seen by a doctor. AI is being used to track, measure, and give significance to a wide range of characteristics, and bodies are, again, thought to possess hidden information which AI can unlock. For instance, faces, voices, emotions (Affectiva, 2018, 2021), brainwaves (Unruly, 2018), morals, honesty (The UK Parliament, 2018), personalities, gender, sexual orientation (Wang & Kosinski, 2018), and political orientation (Kosinski, 2021) are some of the data processed by organisations often using AI systems based on questionable scientific research and practices.

Algorithms can be said to be "the biggest experiment of classification in human history" (Crawford, 2017), or, echoing O'Neil (2016), genuine weapons of mass classification. Furthermore, as noted by Winfield and Karachalios (The UK Parliament AI Committee, 2017c), the invisibility of some AI systems amplifies the unique potential of AI to deceive and create attachment. Thus, the relations between humans and AI powered devices have the potential to become extremely complicated. This is especially the case with Machine Learning (ML), the most successful AI technology (Deng & Yu, 2014; Lecun, Bengio, & Hinton, 2015; Sejnowski, 2018).

Challenging issues are discussed by academics and practitioners and the awareness of discrimination resulting from biased data is growing. Biased data and algorithms leading to discriminatory decisions, lack of transparency in opaque algorithms, and the attempts to use AI to manipulate public discourse are some of the issues causing concern (Cadwalladr & Graham-Harrison, 2018; Greenfield, 2018; Isaak & Hanna, 2018; Tufekci, 2017, 2018).

The debates around AI ethics, and fairness, accountability, and transparency (FAT) of the algorithms are lively and complex (ACM FAccT Conference, 2021; Cyberlaw Clinic Berkman Klein - Harvard, 2019; FAT Conference, 2018; High-Level Expert Group on AI - European Commission, 2019; IEEE Board, 2019; Microsoft, 2019; Organisation for Economic Co-operation and Development (OECD), 2019). The focus is mainly on data and algorithms. Other elements are generally missing from the debate, for example, the praxis of AI and FAT inside organisations. Discussing AI ethics without considering what is happening within organisations, and how these are mediating between different trade-offs while making decisions, is a limitation (from a research point of view) and deeply unfair for those willing to use the technology for good (from an ethical point of view). This is why this research is concerned with how AI and DP are practiced in innovation management.

1.2.2 Data protection, GDPR, FAT and innovation management

Data protection (DP) legislation around the world has gone through major changes in the last few years. DP is particularly strong in the European legislative regime, and the GDPR is the most important legislation. It standardised legislative regimes within the EU, increased the protection of personal data and obligations for organisations, and created new requirements for automated processing. The Regulation has further influenced legislations in other countries, such as the California Consumer Privacy Act of 2018 (California Consumer Privacy Act of 2018 CCPA, 2018), the General Law for the Protection of Personal Data (LGPD) in Brazil (Raul, 2018), and the Washington Privacy Bill (Cesaratto, 2019). The Data Protection Act 2018 (Data Protection Act 2018/DPA) was created in the UK to detail some requirements of the GDPR, and its revised post-Brexit version, the UK GDPR, retains all the GDPR key principles, rights, and obligations.

The FAT principles are also an essential part of the Regulation, as each act of processing personal data must be fair, accountable, and transparent. Data must be processed fairly and transparently, data subjects must be informed on how their data will be used, and organisations must demonstrate compliance with the requirements. This can be challenging for organisations using AI/ML to process data, especially when the processing is carried out with Deep Learning (DP), a type of ML often opaque even to their developers (Koene et al., 2019).

The success of the GDPR is often questioned by those who consider it to be a limit to business entrepreneurship. Digital innovation is creating increasing amounts of data, new business models based on data are replacing old ones, and the digital economy is booming, particularly in the UK, home of several prestigious AI companies, such as DeepMind.

The new "virtual economy", based more on distribution and sharing of information and less on production (Arthur, 2017), is now a consolidated reality. Traditional domains are also merging, requiring new innovation strategies encompassing knowledge of different sectors (The Boston Consulting Group, 2018).

The pace of digital innovation is fast, and the pressure coming from the market increases the need to respond rapidly, often implying very high risks. Embracing digital innovation becomes a necessity for those organisations wanting to guarantee their business continuity. This is usually done via integrating new technology into existing structures or creating new business models.

The development of AI is largely due to ML, the most successful AI technology that has advanced the most in the last few years, and whose breakthroughs (particularly those achieved via DP) are considered to be revolutionary (Deng & Yu, 2014; Lecun, Bengio, & Hinton, 2015; Sejnowski, 2018). Based on predictions, or the guessing of missing information, the use of ML implies a different logic, "prediction instead of rule-based logic" (Agrawal et al., 2018, p. 37), and often a choice between different trade-offs (e.g., more data, less privacy). This also entails different competencies in understanding, creating, and driving innovation, and different approaches to the management of AI (Deloitte, 2018; Luca et al., 2016).

The human factor is crucial in AI. Leaders face new challenges, dilemmas, choices, and new decisions. Decisions are continuously made around AI strategy, AI projects, Autonomous AI (making decisions) and Augmented AI (supporting human decisions). Such choices and decisions do not happen in a vacuum. They are made in context within which specific values and power dynamics operate, both externally and internally to organisations.

1.3 Motivation

Understanding how to manage AI and DP responsibly is imperative considering the pace of AI, and the risks it entails for societies, human rights, and organisations. Research on the relationship between AI, DP, and FAT is just beginning, and it is mainly focused on AI development, developers, data, algorithms, and technical aspects.

Some researchers and practitioners have also started to look at the post-implementation phase of AI systems (Goodman & Flaxman, 2016a), and at the consequences of the technology, such as remedies and mechanisms for redressing violations of rights resulting from decisions made by AI systems. And yet, what occurs in the middle, when organisations choose, implement and use AI, is generally under-researched. AI deployment and use by organisations, the role of stakeholders other than developers, context specificities, dynamics of power, and processes, are still under-explored. Research on "preventive" DP, or the question of when DP strategies are defined, or how and when data protection impact assessments/DPIAs are performed, is scarce. Such an empirical approach to exploring the praxis of AI and DP, and how decisions are made within organisations, is urgently needed. This is especially important considering the growing use of opaque ML making autonomous decisions or supporting human decisions.

This research aims to provide some useful insights into such practices, and by focusing on ML, aims to explore how leaders, managers, and Subject Matter Experts (SMEs) perceive, understand, and apply AI, DP, and the FAT principles, how this affects organisations in the present, and how it is likely to affect their future. In this research, organisations and the people working within them are seen as forces impacting on the innovation created in such spaces. People are not seen as powerless in such contexts. This is especially the case in digital businesses (Griffiths et al., 2018), which are less rigid than traditional ones. Furthermore, due to its speed, AI is more disruptive than traditional technologies (both internally and externally). Thus, digital businesses are more fluid, and have different power dynamics, due to the velocity and frequent disruption that shape the digital economy. Similarly, many GDPR requirements are dependent on practices, communities, and technologies in different contexts (Veale, Van Kleek, et al., 2018). Therefore, this research also considers the relationship between AI and DP praxis and the factors present in a context (e.g., values and stakeholders).

Some frameworks, such as responsible research and innovation (RRI) (more in 1.5), aim at increasing responsible and ethical innovation created by organisations. Similar approaches often face some difficulties. For example, the applicability of RRI to industrial contexts characterised by certain amounts of uncertainty and ambiguity, often intensified by organisational practices.

1.4 Research question, aims, and objectives

This research seeks to answer the following research question:

How can DP and the FAT Principles be applied by organisations during the introduction and use of AI systems and in their digital innovation strategies?

To answer this question, the researcher seeks to achieve the following:

Project Aims

1. Understand the relationship between AI and DP and how they can inform each other in the context of legislation and digital innovation.
2. Examine the extent to which individuals who are adopting and using AI, and DP roles, understand AI, DP, and FAT principles.
3. Understand the impact of DP on organisations that are adopting/using AI, and vice versa.
4. Produce guidance for organisations to support the application of FAT Principles in their AI&DP Management.

Project Objectives

1. To identify how DP legislation protects personal data when processed by AI.
2. To investigate the level of understanding amongst AI adopters and users[1] and DP roles, specifically:
    a. Their knowledge, interpretations and perceptions of AI, DP, and FAT principles.

[1] Investigating end-users is outside the scope of this research.

b. Whether the FAT principles are taken into consideration when AI systems are chosen, implemented, and used.

c. How they use personal data, how they plan to use it, and the current and potential future impact of this on their organisations.

3. To develop a critical theoretical framework that permits the unveiling of the innovation environment and to produce a model aimed at supporting organisations in their AI&DP Management.

The research question, aims, and objectives were developed in the course of the research. They were refined following the internal interim reports and in light of the evolution of the investigation.

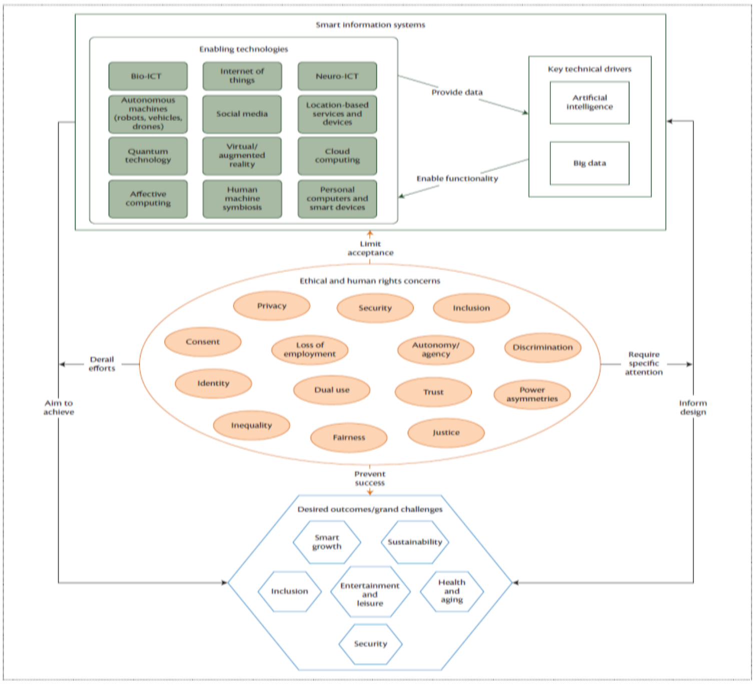
1.5 Theoretical framework

The theoretical framework for this research is an interdisciplinary two-dimensional approach to reading critically the complex relationship between DP, AI, organisations, and people.

Drawing upon responsible research and innovation (RRI) (European Commission, 2020; Orbit, 2020; Owen, 2014; Owen et al., 2012; RRI Tools, 2020; B. C. Stahl, 2018; B. C. Stahl, 2012; B. C. Stahl et al., 2017), and critical theory of technology (CTT), and specifically the work by Feenberg (1991, 2002, 2005), the framework adopts the holistic approach needed to read the complexity of organisational practices.

1. RRI refers to a framework for European Programmes (European Commission, 2020) that takes into consideration the impact of research and innovation on societies. It involves stakeholders since the beginning of the process and considers consequences and social and moral values. It is “a collective, inclusive and system-wide approach“ (van den Hoven, 2013, p. 3), which considers various elements and ethical concerns (Figure 1-1) and can offer a framework for ensuring that the technologies are “socially acceptable, desirable and sustainable” (B. C. Stahl & Wright, 2018a, p. 1).

Figure 1-1 The ecosystem of smart information systems



(Source: Stahl & Wright, 2018, p 28)

However, RRI has some limits. It is mainly related to large-scale research projects carried out in the public sector (ibidem) and not to innovation created in the private sector, whose dynamics, concerns, and pace are very different. It can also be interpreted as a top-down framework for ideal situations (mainly focused on research innovation and less on deployment and business), which does not consider specific or practical issues (e.g., different organisational contexts and power imbalances between different stakeholders). Despite these limitations, RRI is considered to be an extremely valuable approach to explore complex contexts.

2. Feenberg, and CTT, were influenced by the Frankfurt School, Heidegger, and social constructivism. Technology is seen as being socially shaped (Zheng & Stahl, 2011), it is not considered to be neutral and is subject to power relations. This social perspective is particularly useful for this research. Feenberg's analysis of technologies as instrumental in creating modern hegemonies (Feenberg, 2002) is of relevance for considering accountability, responsibility and power in societies increasingly regulated by algorithmic systems. He does not believe in the "occult power of the 'technical phenomenon', rather, technology, as a domain of perfected instruments for achieving well-being, is simply a more powerful and persuasive alternative than any ideological commitment" (ibidem, p. 12). Technology is seen as "'an ambivalent'

process of development suspended between different possibilities" (ibidem, p. 15), an ambivalence which, again, does not imply neutrality. Technology is therefore ambiguous and is a mirror of the values existing in the space where it is developed or adopted.

Furthermore, this research demands the adaptation of a critical theory of subjectivity which explores limitations with regards to people's agency. As it is people within organisations and systemic settings who adopt and develop these technologies, the question of responsibility exceeds the individual dimension. Therefore, this research draws upon some specific elements of RRI and CTT, which help to read the praxis of AI, DP, and FAT, and explore the complexity of organisational contexts. This theoretical framework encompasses various elements, such as:

- Societal values, and how they are understood and perceived.
- Context and culture.
- Subjects' experience and influence on praxis.
- Risks, processes, and stakeholders.
- Decision-making processes.
- Power and power dynamics.

Therefore, a combination of RRI and CTT can help to identify how the technology is implemented, how the decisions are made, and under which conditions and assumptions.

1.6 Impact

The research provides critical knowledge on a highly topical issue. Organisations are investing in GDPR compliance, often unaware of or underestimating its complexity and the risks of non-compliance. Many are choosing AI, often as a quick fix, risking the disruption of internal and external equilibria. At the same time, the growing debate on AI ethics is mainly based on general principles that often lack the point of view of those implementing innovation within organisations. This research fills this gap by:

a. Providing insights on the intersection of AI, GDPR and FAT.

b. Offering tools to help organisations understand the challenges, risks, and benefits.

c. Supporting organisations in using AI in a fair, transparent, and accountable way, and helping them to foster new responsible management.

1.7 Contribution

This research addresses the management of AI in practice. It identifies a gap in the existing theory, defines key elements within organisational contexts, and creates a distinctive framework for operationalising AI/ML that can be used by leaders and managers in their innovation management.

Additionally, this research looks at the current discourse of FAT Principles and its relation to AI ethics. While the debate on AI ethics is growing, this has been explored mainly in relation to technical aspects. Other elements are still under-explored, e.g., the role of those implementing AI and their understanding of fairness.

The research provides practical guidance on how FAT principles can be applied in AI management, and how this can be developed further by organisations interested in enhancing practical responsible management.

Specifically, this research provides a significant original contribution to:

- Theory - the research elaborates critical AI&DP/*CRAIDA*, a new theoretical management framework, which addresses the management of the technology in practice.
- Knowledge - the research creates responsible IS AI&DP/*RAIDIS,* a management model for organisations adopting responsible AI management.
- Practice - the research creates *RAIDIS* maturity model/*RAIDIS MM*, a model for organisations adopting responsible AI management, and *Responsible Augmented AI* that unveils the complexity of decision making in Augmented AI models.

1.8 Research methods

This study uses an interpretivist and inductive approach and qualitative methods as its methodological choice. This research methodology was developed to address the aims and objectives of the research. Semi-structured interviews and case studies (multiple case approach and methods) (Table 1-1) were adopted to understand the relationship between AI, DP, and FAT, and to provide answers to complex contemporary phenomena (what, how, and why questions) via the analysis of beliefs, interactions, and experiences occurring within organisational contexts.

Table 1-1 Data collection methods

Data Collection Methods		
Expert insights survey (Interviews)	**Case Studies**	
	CS 1	**CS 2**

An expert insights survey was the first research step. Nine participants (experts in ML, business technology, DP, and privacy) provided information from different sectors and from their assignments in many organisations. That was followed by two cases studies involving another 11 participants (leaders and senior managers, and ML and DP roles). Case studies were considered particularly appropriate for this research. Widely used in Information Systems to understand contemporary phenomena, they offered multiple data sources (interviews, document analysis, and observations), in-depth insights of the settings, and extensive understanding of organisational practices. Organisations were based in the UK and were chosen among those that are planning, implementing, or already using AI/ML. Details on the case study design are provided in Chapter 3.

The analysis of the data was conducted using a thematic analysis approach that permitted the identification, analysis, and reporting of patterns (themes) in the data. Coding was performed using the software NVivo.

The results offer a detailed understanding of organisational practices, individual experiences, and the implications faced by organisational implementing AI/ML and the GDPR.

1.9 Definitions of terms

Some terms are used throughout the thesis. Their specific definitions in the context of this research is provided below:

-<u>Artificial Intelligence (AI) and Machine Learning (ML).</u> AI is used when the text refers to AI technologies in general, inclusive of all technologies developed within the realm of the AI tradition. ML is used when the text specifically refers to ML, the most successful subset of AI, which learns from experience, improves, and makes predictions.

-Data protection (DP) and the GDPR/Regulation. DP is used when the text refers to DP legislation in general. GDPR is used when the text specifically refers to the Regulation, its requirements, and articles.

-AI management. In the context of this research, AI management refers to both strategic and operational management of AI. The management of AI is viewed not only as the management of the technology, independent from other aspects of organisational management, but is understood as a holistic management inclusive of other factors present in the context (e.g., strategy or management of stakeholders). Additionally, the use of AI to manage staff is outside of scope of this research.

-Organisations. This research refers to entities encompassing people, technology, processes, and stakeholders which operate within the public/private sector.

-Digital Innovation. The application of digital technology to business problems. This thesis is in particular concerned with the use of AI in digital innovation.

-Framework. A structure that provides the support for a system or a set of concepts.

-Model. A practical tool that describes how different parts of an organisation can work together successfully.

A more detailed list of terms and definitions used in this thesis can be found in the Glossary.

1.10 Thesis outline

This thesis is organised into nine chapters.

This introductory chapter has introduced the context, motivation, impact, and contribution of the research. It has also presented the research question and the project's aims and objectives, theoretical framework, research methods and some definitions of terms.

Chapter 2 presents the Literature Review. The first part discusses AI, its evolution, and some current applications, with special focus on ML. The second part focuses on data protection, the evolution of Human Rights in the EU legislation, the GDPR, the Data Protection Act 2018, the UK GDPR. The third part analyses the FAT Principles, while the fourth introduces AI management, traditional and digital business, and focuses on the specificities of ML. Chapter 3 illustrates the research methodology used for this research. Chapters 4, 5 and 6 present the findings from the interviews with the group of experts, Case Study 1 (CS1) and

Case Study 2 (CS2). Chapter 7 focuses on the new *CRAIDA* theoretical framework used in the research, and Chapter 8 presents the *RAIDIS* models, practical and strong mechanisms for governing complex innovative environments that identify exact risks and potential within specific areas. Chapter 9 ends with the conclusion, highlighting the need for a holistic approach to the management of AI, and the key role of subjects in defining the responsible use of the technology. The chapter then finishes by making suggestions for future research (i.e., the management of Augmented AI).

CHAPTER 2: LITERATURE REVIEW – AI, DATA PROTECTION, FAT PRINCIPLES, AI MANAGEMENT

2.1 Introduction

This chapter reviews the literature on AI, DP, and the FAT principles, and examines digital innovation management and the emerging AI management.

The review begins with a focus on the concept and evolution of AI, followed by an explanation of different classifications, areas and approaches, and main current and forecasted applications, with a special focus on the implications for DP and privacy. The second part focuses on DP, the evolution of the right to DP in Europe, GDPR, Data Protection Act 2018 and the UK GDPR, their significance for AI systems, and the current debate on FAT and biases. The chapter ends with a presentation of the main characteristics of digital innovation management and the differences from traditional business, focusing on AI management, main AI strategies, and its specific challenges in managing ML.

2.2 Artificial Intelligence

2.2.1 What is AI?

Breakthroughs and rapid developments in AI and in ML in the last few years have been followed by a growing debate on AI, its potential, applications, and challenges. AI is both inspiring and concerning. Already deployed in many systems, AI is no longer confined to the domain of Computer Science. Its meaning and current and future applications are being discussed by experts of various disciplines and by individuals who are becoming more aware, curious, and fearful. And yet, a great amount of confusion still surrounds the significance of AI, its various classifications, types, and technologies.

There is not a unique definition for AI. Various concepts are used to describe it, as various approaches exist to understand intelligence, consciousness, and the relationship between the mind and the body. The mystery around AI contributes to the debate in various areas, for example, around Human-AI interaction, ethics, and morals (e.g., regulations, or a morality for a future Artificial General Intelligence), and philosophy and psychology (especially in relation to questions related to the mind, mental states and consciousness of AI systems).

AI is defined, in general, as the capacity of a machine to perform or think like a human being, in opposition to Natural Intelligence (NI) which belongs to biological systems. Norvig and Russell (2016) classify various definitions of AI into human approach and rational approach:

1. Human approach: computers think or act like humans, and the more similar they are, the more successful AI systems are (Bellman, 1978; Haugeland, 1989; Kurzweil et al., 1990; Rich & Knight, 1991). Computers carry out mental processes (cognitive modelling approach) or act like human beings. For example: Haugeland believes that the ability of a human to elaborate thoughts (as a symbolic representation of reality) is "radically the same" as that of a machine (p. 2), while Bellman conceptualises AI as the capacity to acquire new knowledge, and to use it critically to make decisions and to solve problems. Kurzweil et al (1990) and Rich and Knight (1991) see AI systems more as being able to act rather than to think like humans: "Artificial Intelligence (AI) is the study of how to make computers do things which, at the moment, people do better" (Rich & Knight, p. 3).
2. Rational Approach: Computers and machines perceive, reason and act logically like a rational agent (Poole et al., 1998; Winston, 1992).

The lack of consensus on a specific definition is also cited by Calo (2017). AI is, in general, viewed as a complex of techniques aimed at simulating perception and reasoning similar to biological beings. He highlights the difficulty of creating a technology capable of performing all cognitive tasks (as forecasted in the '50s and '60s): "(w)hat seems possible in theory has yet to yield many viable applications in practice" (ibidem, p. 3).

2.2.2 History of AI

AI has advanced in different areas over the last few years, its breakthroughs have stimulated debate, and it is no longer just a feature of fiction or computer science.

The idea of an enhanced humanity resultant of a superior technological knowledge is not new. The desire to create entities or machines able to think and act like humans is present in the work of various authors. From the character Pygmalion created by Ovid in the Metamorphoses (McConnell, 2007; Ovid, 1968), or the myth of the criminal humanoid Frankenstein (Lehman-Wilzig, 1981), to the films Terminator, Matrix, Her, or the dystopian Blade Runner, the idea of an enhanced humanity resultant of a superior technological knowledge has always been part of the (Western) popular culture (Schmerheim, 2018).

However, modern AI began in the ’50s with the work by Alan Turing and his “Computing Machinery and Intelligence” (1950), and the Dartmouth Conference in 1956.

Turing started presenting his ideas in his lectures from 1947 onwards (Norvig & Russell, 2016) and created and presented the Turing test in his seminal paper “Computing Machinery and Intelligence” (1950), considered by many as the starting point of modern Artificial Intelligence and a milestone in the history of computers (Finlay & Dix, 1996). In the paper Turing presents his famous “imitation game”:

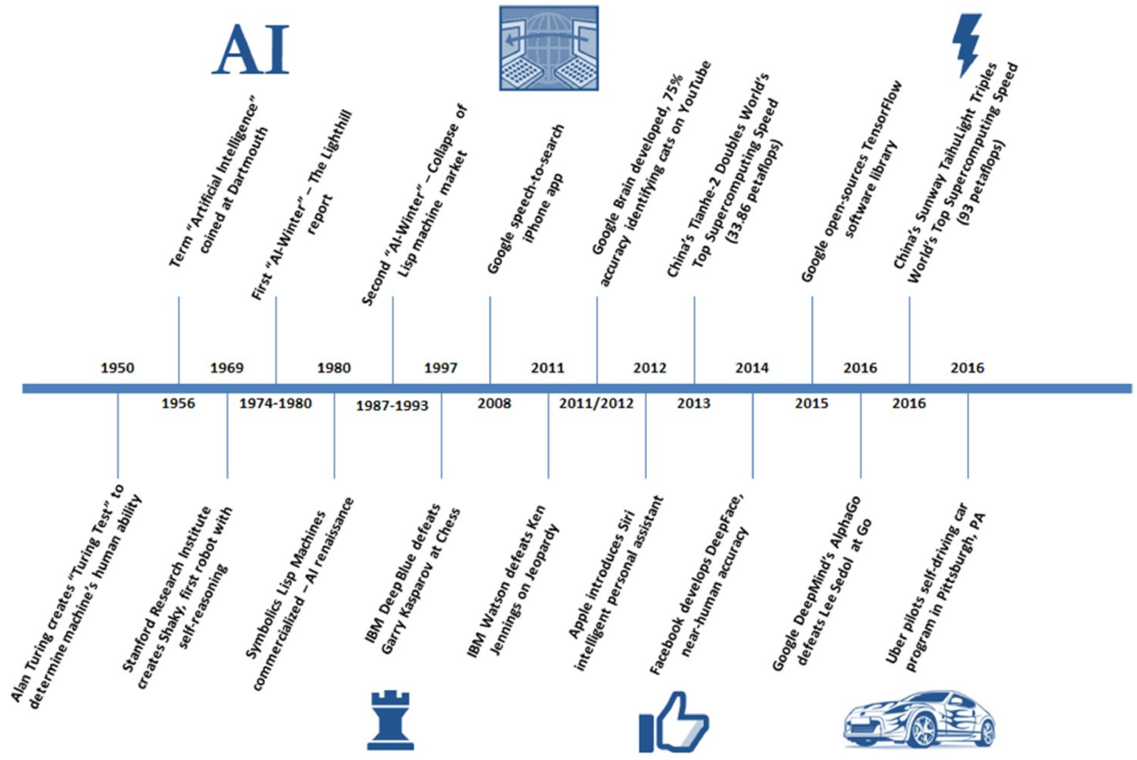
> I propose to consider the question, “Can machines think?” This should begin with definitions of the meaning of the terms “machine” and “think.” The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words “machine” and “think” are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, “Can machines think?” is to be sought in a statistical survey such as a Gallup poll. But this is absurd…(Turing, 1950, p. 1).

As the definitions of the terms “machine” and “think” can be unsatisfactory, he suggested an experiment (the Imitation Game), in which the machine is said to “have passed” the intelligence test if the player, in their interaction with both the machine and another human, is not able to distinguish one from another. Turing predicted that a machine would have been able to pass the test by the end of the century. The test was said to have been passed by the chatterbot Eugene Goostman in 2014, but this was met with scepticism (Sample & Hern, 2014; Warwick & Shah, 2016). The more recent development of more sophisticated conversational AI agents, able to make people believe they are interacting with human beings, has produced conflicting opinions on their capacity to pass the test, furthermore, raising concerns on the ethics of some systems, potentially able to deceive, manipulate and impersonate real people, as can be the case with chatbots (Brundage et al., 2018) and other generative models (Zhang et al., 2021).

The name Artificial Intelligence was created by John McCarthy in 1955 (McCarthy et al., 2006), and entered into mainstream research as a new discipline during the Dartmouth Summer Research conference on AI organised by McCarthy in 1956.

Figure 2-1 Evolution of AI to 2016

Exhibit 8: Evolution of AI: 1950-Present


AI
Term "Artificial Intelligence" coined at Dartmouth
First "AI-Winter" – The Lighthill report
Second "AI-Winter" – Collapse of Lisp machine market
Google speech-to-search iPhone app
Google Brain developed, 75% accuracy identifying cats on YouTube
China's Tianhe-2 Doubles World's Top Supercomputing Speed (33.86 petaflops)
Google open-sources TensorFlow software library
China's Sunway TaihuLight Triples World's Top Supercomputing Speed (93 petaflops)
1950
1969
1980
1997
2011
2012
2014
2016
2016
1956
1974-1980
1987-1993
2008
2011/2012
2013
2015
2016
Alan Turing creates "Turing Test" to determine machine's human ability
Stanford Research Institute creates Shaky, first robot with self-reasoning
Symbolics Lisp Machines commercialized – AI renaissance
IBM Deep Blue defeats Garry Kasparov at Chess
IBM Watson defeats Ken Jennings on Jeopardy
Apple introduces Siri intelligent personal assistant
Facebook develops DeepFace, near-human accuracy
Google DeepMind's AlphaGo defeats Lee Sedol at Go
Uber pilots self-driving car program in Pittsburgh, PA

Source: Company data, Goldman Sachs Global Investment Research

(Source: Goldman Sachs, 2016)

The conference now considered the formal beginning of AI (see Figure 2-1) was also the beginning of research on AI carried out at some important U.S. universities, such as MIT and Stanford, with funding provided by the Defence Advanced Research Project Agency (DARPA) (Chiou et al., 2001).

The conference was not only the beginning of AI, but it was also the beginning of its demographic specificity, which has shaped and defined the AI research environment since then. The AI community has been a predominantly white, male-dominated research environment and a narrowly defined community (Campolo et al., 2017), a very specific characteristic of AI ecosystems that will be investigated later while analysing FAT in AI (see 2.4).

Another important moment in the history of AI was the creation of Eliza (Weizenbaum, 1976), an early version of a Natural Language Processing programme which can be considered the first chatbot. The programme used a patter matching or string-matching algorithm to match some common expressions and using some specific scripts appeared to understand the meaning and the context of the communication. Particularly famous was a script called the Doctor, in which the computer acted as a therapist that used some counselling tools, such as open-ended questions and reflection techniques. The programme was in some cases so successful that some individuals became emotionally attached to it, an outcome unexpected and unwelcomed by Weizenbaum.

Considering the current debates on AI ethics, some of Weizenbaum's considerations about the danger of AI, the morality of the AI research environment, and the deceptive power of AI are particularly interesting. Weizenbaum believed AI to have certain limits, especially in relation to human creativity, and he was highly sceptical of the "most extreme fantasies of the artificial intelligence community, 'the artificial intelligentsia', hence irrelevant to practical current concerns" (Weizenbaum, 1978, p. 14): a direct reference to the perception and understanding of knowledge and facts. Similar issues are still being debated in the AI ethics discourse, for example, in relation to the potential deceptive power of AI systems (Brundage et al., 2018; High-Level Expert Group on AI - European Commission, 2019; Gershgorn, 2017; Goodfellow et al., 2017), or the power of fake news, false information distributed mainly via social media for political propaganda (Allcott & Gentzkow, 2017), or Deep Fakes, fake videos and audio clips created with Deep Learning (DP) (Albahar & Almalki, 2019; Chesney & Citron, 2018).

The interest in AI has gone through various periods since the '70s: intense curiosity, hopes and investment (AI Spring), but also reduced interest and resources (AI winters) (N. J. Nilsson, 2009) when AI failed to deliver the expected progress. A new AI renaissance started in the '90s, thanks to enhanced computer power, amount of data, and technical capabilities. For example, IBM's T.J. Watson Research Center created Deep Blue (1997), a chess software that won against the world champion Garry Kasparov (Campbell et al., 2002), and Watson (2011) a "question answering machine" (Markoff, 2011, p. 1) that beat two champions and won the TV game show "Jeopardy!".

Even though some of the initial AI hype receded (Di Ieva, 2019; Floridi, 2020; The Economist, 2018a), breakthroughs, investments and interest in AI as a strategic means for economic, political and geopolitical gain will escalate its use (Center for AI and Digital Policy, 2020; Future of Life Institute, 2017; Knight, 2016; Zhang et al., 2021).

2.2.3 Narrow AI and Artificial General Intelligence (AGI)

The new renaissance has been the result of a significant change occurring at a theoretical level, when the focus shifted from AI to ML, and the debate moved from Good Old-Fashioned AI (GOFAI) to Narrow AI and the future Artificial General Intelligence (AGI).

- Good Old-Fashioned AI (GOFAI) aimed at creating a general intelligence based on symbols, similar to human intelligence and capable of understanding various problems. After acknowledging the limitations of practical applications, the approach of the scientific community moved on to specific areas/subfield of AI.

- Weak or Narrow AI aims at solving issues in specific domains, for example playing chess, by using methods and models from other fields, such as statistical and pattern-recognition approaches (Langley, 2011).

  As highlighted by Harvey (2000), this shift reflects the embracing of a different philosophical tradition, from a Cartesian and classic approach, with only one version of the reality (and intelligence) which is not dependent on the observer, to a more subjective perception of reality. Intelligence is seen as a "form of adaptive behaviour amongst many" (ibidem, p. 15), and cognition is understood as "the priority of lived phenomenal experience, the priority of everyday practical know-how over reflective rational knowing-that" (ibidem).

- Artificial General Intelligence (AGI), or Superintelligence or Singularity, is a potentially superior version of AI capable of self-evolving and of surpassing human intelligence. The timeframe and consequences for humanity are unclear, as researchers and AI practitioners hold different opinions on outcomes and existential risks (Calo, 2017, p. 4). Notable is the work on AI conducted at the Machine Intelligence Research Institute (Berkeley, 2019), and at the Future of Humanity Institute (University of Oxford - Future of Humanity Institute, 2019).

  While AGI is expected to become more relevant for future organisations, raising questions of regulation and public policy, it is beyond the scope of this project.

  Various projects are working on AGI (Baum, 2017), such as Open AI, which recently released Generative Pre-trained Transformer 3 (GPT-3), a powerful language prediction model (not immune from reproducing Islamophobic content), and DeepMind, whose mission is to create AGI, the most successful Deep Learning (DL) company.

Over the last few years, the major breakthroughs have been achieved using DL, the area of ML based on Artificial Neural Networks (NNs) (The Economist Science and Technology, 2018). By allowing “multiple processing layers to learn representations of data with multiple levels of abstraction” (Lecun et al., 2015, p. 1), DP simulates the activity of layers of neurons in the brain (Schmidhuber, 2015; The Economist Science and Technology, 2018).

DeepMind combined (for the first time) deep neural networks with Deep Reinforcement Learning (two different approaches in two different stages). The system learned how to play the game GO from human data using stochastic searches and deep neural networks. It then improved its skills by playing against itself by using reinforcement learning (The Royal Society, 2017), making the human experience redundant (Silver, Hubert, et al, 2017). DeepMind aims at using less data (and personal data) in training their algorithms (Hassabis & Silver, 2017).

Owned by Alphabet (DeepMind, 2021), DeepMind’s breakthroughs are milestones in the AI evolution, and particularly promising is the research by DeepMind Health, the division applying ML to healthcare (Harwich & Laycock, 2018, Venkataramakrishnan, 2020). And yet, some partnerships it entered into in order to obtain the necessary data to train its algorithms have raised many questions around DP (Powles & Hodson, 2017; Shead, 2017), and have prompted an investigation conducted by the Information Commissioner’s Office (Denham, 2017; The National Data Guardian, 2017). Furthermore, some of the issues with Google, which created its own health-care division rolling out DeepMind Health under its direct control (Hodson, 2019), raise further questions about data ownership in the case of company mergers and acquisitions.

2.2.4 AI areas

AI is an umbrella term which comprises different scientific technologies aimed at solving specific problems through the reproduction of biological processes, similarly to human cognitive abilities (see Figure 2-2). In the common imagination, central to the discipline of AI is the idea of autonomous systems that act independently of individuals. However, AI is also used to add “knowledge and reasoning to existing applications” (Mata et al., 2018, p. 1).

AI technologies are mainly grouped by: Search and Problem Solving; Knowledge, Reasoning and Planning; Communication and Speech, Movement, Learning.

a. Search and Problem Solving: The ability of the AI system to search for the best solution to a specific problem, starting from a situation where not all information is known. Example of tools are search trees and data structures (i.e., number of nodes) which consider various factors such as constraints related to the amount of time and memory (i.e., search algorithms or evolutionary computation).

b. Knowledge, Reasoning and Planning: Knowledge of the world, perception and representation according to various systems of values, and then the action following the analysis of that information:

   - Knowledge Representation is the power to represent knowledge, "the most important aspects of the real world, such as action, space, time, thoughts, and shopping" (Norvig & Russell, 2016, p. 437). The organisation of knowledge into categories, subcategories, subclasses, and their relations (ibidem, p. 441) is for example a characteristic of knowledge representation.

   - Reasoning: the capacity of AI to solve problems through logical deduction via available knowledge (i.e., puzzles or games).

   - Planning: the capacity to devise a plan of actions and reach those planned goals. An ideal future status is predicted, and planned actions are set up in order to reach it.

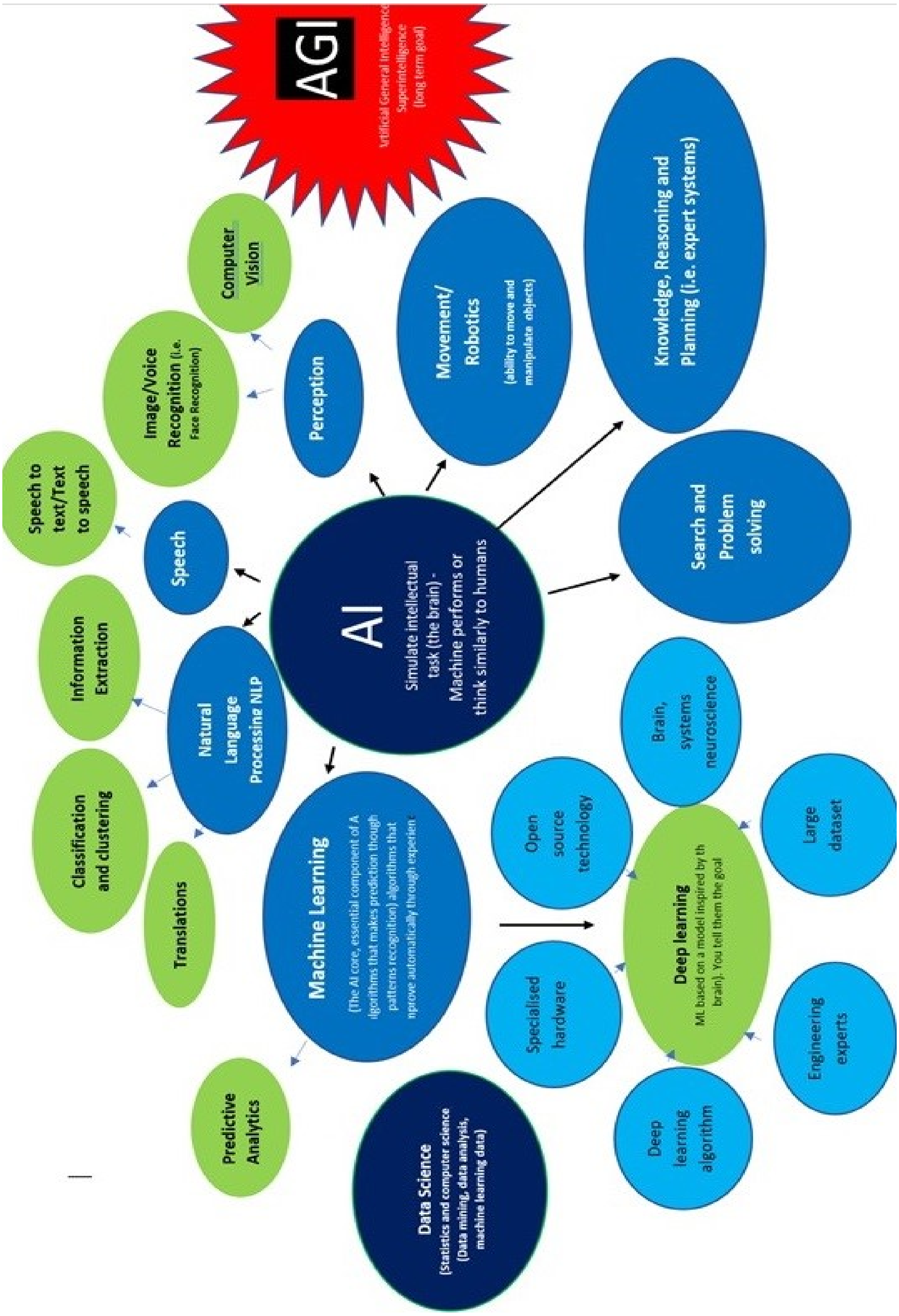
c. Communication and Speech: the ability to understand languages (written and spoken), to create language (Brown et al., 2020) via Natural Language Processing and to create text from speech and vice versa.

d. Perception: the capacity to perceive the external world via sensors and to recognise images, videos and sounds (such as voices). Examples of AI technologies are computer vision, voice, image, and face recognition.

e. Movement: the ability of machines/robots to perceive, make decisions, move in different environments, and manipulate objects.

f. Learning: the capacity of some AI systems to learn from experience and examples, improve and make predictions (ML).

Figure 2-2 AI technologies and applications


AGI
rtificial General Intelligenc
Superintelligence
(long term goal)
Knowledge, Reasoning and Planning (i.e. expert systems)
Movement/ Robotics
(ability to move and manipulate objects)
Perception
Computer Vision
Image/Voice Recognition (i.e. Face Recognition)
Speech to text/Text to speech
Speech
Information Extraction
Search and Problem solving
AI
Simulate intellectual task (the brain) - Machine performs or think similarly to humans
Natural Language Processing NLP
Classification and clustering
Translations
Machine Learning
(The AI core, essential component of A lgorithms that makes prediction though patterns recognition) algorithms that nprove automatically through experien
Predictive Analytics
Data Science
(Statistics and computer science (Data mining, data analysis, machine learning data)
Brain, systems neuroscience
Open source technology
Large dataset
Deep learning
ML based on a model inspired by th brain). You tell them the goal
Specialised hardware
Engineering experts
Deep learning algorithm

(Design: Chiara Addis)

2.2.5 ML

This research focuses on ML, the area of AI based on predictive technology that has advanced the most in the last few years, and whose breakthroughs (particularly those achieved via DP) are considered revolutionary (Deng & Yu, 2014; Lecun, Bengio, & Hinton, 2015; Sejnowski, 2018).

The term was created by Arthur Samuel (Samuel, 1959), who first created a ML programme able to learn through playing games. He believed it was possible to teach a machine to learn how to play checkers better than the programmer, by giving it the ability to learn without being explicitly programmed. Using games "provides a convenient vehicle for such study as contrasted with a problem taken from life, since many of the complications of detail are removed" (ibidem, p. 211). As previously seen with IBM and DeepMind, training ML systems by playing games proved successful also with other researchers.

ML is the result of the combination of Computer Science with principles and elements from different fields, notably Statistics and Maths. While traditional programming is based on hard-coded sets of rules, which are rigid and fixed, ML analyses massive amounts of data, identifies patterns and correlations, and makes predictions and, in some cases, decisions.

The success of ML is closely linked to the diffusion of Big Data and Cloud Computing. Big Data is related to this large set of data described with the 3Vs model:

- High Volume of data (not of samples).
- High Velocity (high speed and real-time analysis).
- Variety of information from different sources. Data is extracted, integrated and combined for a deeper analysis, and this is often done with AI-Machine learning technologies) (Diebold, 2012).

More recently, other scholars (Fenton et al., 2019) used the four Vs, volume, velocity, variety, and veracity (accuracy and credibility of the data), to describe how data is growing in a world characterised by volatility, uncertainty, complexity, and ambiguity (VUCA). A fifth V is identified in the value created by the organisation.

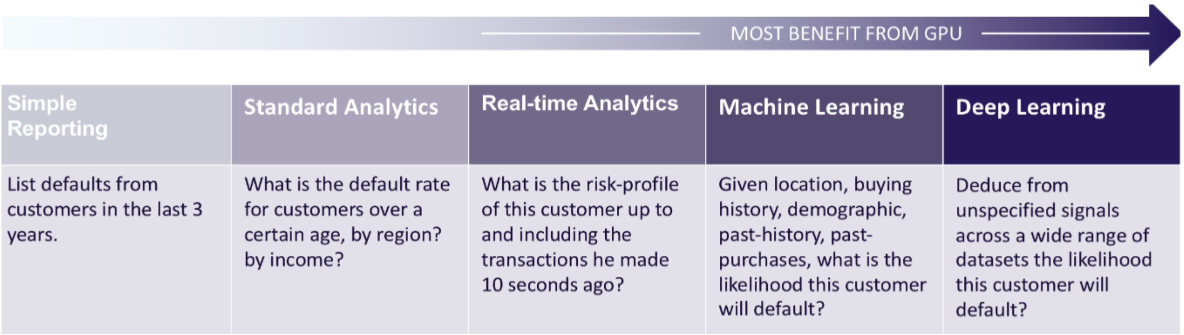
The volume of processed data, the speed of processing and the variety of sources make the difference, and this allows ML technologies to analyse huge amounts of Real-Time Data (RTD) and make predictions and almost real-time decisions. ML can process "volumes of data that would be unmanageable for humans…extracts value by deriving new insights from the mass of data, and in turn data is needed to develop Machine Learning, by training systems to detect patterns or make predictions” (The Royal Society, 2017, p. 49).

The nature of the relation between AI and massive amounts of data was also highlighted by Buttarelli, the late European Data Protection Supervisor, who saw it as bi-directional: “Artificial intelligence, through machine learning, needs a vast amount of data to learn, data in the realm of big data considerations. On the other direction, big data uses artificial intelligence techniques to extract value from big datasets” (European Data Protection Supervisor, 2016, p. 4).

ML facilitates the shift from descriptive to prescriptive. By using huge computer power at a higher scale and speed than simple predictive analytics, the analysis shifts towards shaping future outcomes (see Figure 2-3). Starting from descriptive analytics, that analyse what happened in the past, the analysis evolved towards prescriptive analytics.

Predictive analytics performed with ML predict what could happen, and by providing advice on potential outcomes, provide information that could be used to shape outcomes.

Figure 2-3 Analytics vs ML and DP


MOST BENEFIT FROM GPU

Simple Reporting	Standard Analytics	Real-time Analytics	Machine Learning	Deep Learning
List defaults from customers in the last 3 years.	What is the default rate for customers over a certain age, by region? by income?	What is the risk-profile of this customer up to and including the transactions he made 10 seconds ago?	Given location, buying history, demographic, past-history, past-purchases, what is the likelihood this customer will default?	Deduce from unspecified signals across a wide range of datasets the likelihood this customer will default?

(Source: Dilworth, 2017)

Therefore, prediction is the key element in ML. “PREDICTION is the process of filling in missing information. Prediction takes information you have, often called ‘data’, and uses it to generate information you don’t have" (Agrawal, Gans, & Goldfarb, 2018, p. 24). Thus, predicting from a dataset is not forecasting. ML uses past data to predict, or guess, hidden or missing information, a sort of technological “crystal ball” for guessing past, present, and future (ibidem). Introna (Rowe et al., 2020) expands this, seeing ML as correlation machines and correlation practices which “prioritise prediction over explanation in what is called the post theory paradigm…” (ibidem). Unable to predict the future, they can “create the future they already assumed for their operation to make sense….they conceal their own normativity and own politics behind the façade of neutral calculative practices…often presented as reflecting the world as it is" (ibidem).

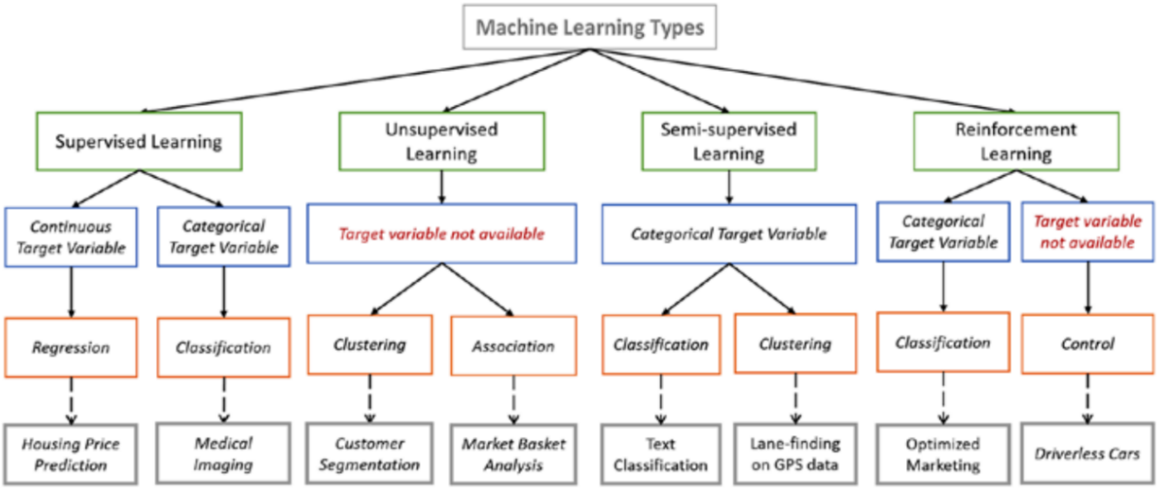
As clearly highlighted by Agrawal, Gans, & Goldfarb, the role of prediction in business strategies is mainly the success of predicting machines. Making a clear distinction between AI and ML, they observe how "the new wave of artificial intelligence does not actually bring us intelligence but instead a critical component of intelligence — *prediction*." (p. 2). The AI technology used in translation, speech to text, identification of illnesses such as cancer, or genetics in DNA analysis for the early identification of potential diseases, prevention of crime and security (Fuster, 2020)*,* and in other similar applications, is based on the prediction generated by ML.

Therefore, applications based on prediction are a precious tool for organisations desirous of reducing costs and improving their efficiency, and they are "a microcosm of what most businesses will be doing in the near future" (Agrawal, Gans, & Goldfarb, 2018, p. 3). As good predictions reduce uncertainty and inform better decisions, implementing a technology able to provide reliable predictions has become particularly important for organisations needing to consolidate their position in the market, especially in moments of high political uncertainty and economic instability.

a. ML classification

Machine Learning is classified into four main types of learning (Figure 2-4):

Figure 2-4 Types of ML



(Source: Ramasubramanian & Singh, 2017)

1. Supervised Leaning: Algorithms are trained using datasets that have been labelled according to groups or categories. The system learns what is included in that group and what is not, and throughout this process it creates rules, or models, used to understand the environment, to adapt to it, and to make predictions.
2. Unsupervised Learning: Algorithms do not learn via predefined categories, but they are set free to find similar characteristics in the dataset. "…[T]he agent learns patterns in the input even though no explicit feedback is supplied. The most common unsupervised learning task is clustering" (Russell and Norvig, 2016, p. 694). For example, the system finds structures, creates clusters and learns from assigning values to them, or "it will seek to determine characteristics that make the data points more or less similar to each other and will attempt to represent the data in a summary form, such as through clusters or common features (The Royal Society, 2017, p. 123).

   Generative adversarial networks (GANs) are one of the most successful applications (I. J. Goodfellow et al., 2014), especially in image recognition, and are based on "adversarial nets framework" (p. 1), where two neural networks, trained on the same data, contest each other (the generator and the discriminator) via an adversarial process which progressively improves their methods.
3. Semi-supervised Learning: Only part of the data sets is labelled, and the system learns from it. Particularly useful in complex experiment, this method "include(s) probabilistic models, graph-based semi-supervised learning, and transductive Support Vector Machines" (ibidem). For example, computer vision and drug tests.
4. Reinforcement Learning: It is based on the interaction of an agent with his environment. Learning is the result of external inputs, continuous interactions, decisions, rewards, learning and adaption. The goal of the agent is to learn the consequences of its decisions, such as which moves were important in winning a game, and to use this learning to find strategies that maximise its rewards (The Royal Society, 2017, p. 20). "Methods in this field include Q-learning, direct-policy methods, and PILCO" (ibidem, p. 123). For examples, games and robots, whose software use RL to make sense of the space in which they move.

b. ML current applications

ML technologies are already being used in several sectors (e.g., transport, healthcare, finance) and in various applications (e.g., personal recommendations, virtual personal assistants, image processing, pattern recognition and anomaly detection).

-Recommendation of products or services. ML systems find patterns in user and other individuals' past preferences and make suggestions (predictions) accordingly. Examples include the technology used by Facebook, Amazon, Alibaba, Netflix, Google.

The same technology can be used to manipulate public opinion, targeting minority groups (Angwin et al., 2017), and suggesting specific content ahead of political elections (Tufekci, 2017, 2018). Its use by Cambridge Analytica without individuals 'awareness or consent (Cadwalladr, 2017; Greenfield, 2018) is an example of the potential malicious use of AI (Brundage et al., 2018), and this was predicted by Kosinski, Stillwell and Graepel in their paper (Kosinski et al., 2013) describing the idea for the model later used by Cambridge Analytica. Predictions can have major and negative implications for the individual's privacy and safety:

> (T)he predictability…can easily be applied to large numbers of people without obtaining their individual consent and without them noticing. Commercial companies, governmental institutions, or even one's Facebook friends could use software to infer attributes such as intelligence, sexual orientation, or political views that an individual may not have intended to share (ibidem, 2013, p. 4).

Another consequence is the creation of "filter bubbles", a phenomenon able to impact on people isolating them from diverse viewpoints and experience (Nguyen et al., 2014). Through personalised recommendation systems, algorithms predict what the user might want to see and filter content according to users 'preferences, de facto "…limiting people's experience and trapping them in echo chambers in which they find their existing views and prejudices reinforced and amplified" (House of Lords, 2018, p. 331). The potential negative effects are amplified by their pervasive use in different systems which mediate experiences of users who are not always aware, and by their capacity to contribute to the polarisation of socio-political opinion and manipulation of public discourse (Seargeant & Tagg, 2019).

Therefore, filter bubbles have the potential to reinforce existing views, alter the perception of free choice, limit the desirability and the experience of otherness, and impact socio-political orders.

-Virtual Personal Assistant, Speech to Text and Voice Cloning. Amazon Alexa and Google Home are the most popular assistants, and the best example of Internet of Things (IoT) technology, which connects devices in a space (e.g., home, city) and is able to continuously collect, share and send information. Virtual Personal Assistants can constitute a real risk for personal data (Monitor Research Group, 2018).

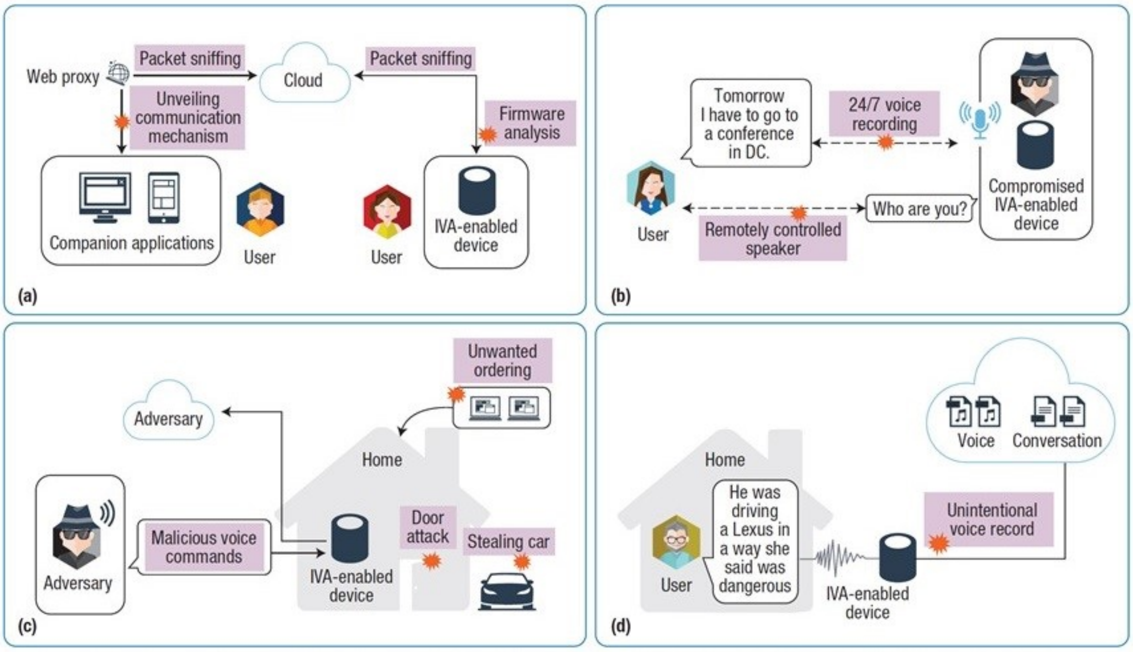
News of devices sending personal information of people directly from their houses is now more common, with some instances raising concern. Some recent cases reported by Media and Researchers have serious implication for privacy and DP of users:

- Some devices were reported recording and sending conversations to people on their contact lists (Chokshi, 2018).
- Unlawful recordings of millions of children made to save their voice prints, leading to a class action presented on behalf of children in eight US states (Kaitlyn, 2019).
- A global team of Amazon employees was reported to have been listening and reviewing users' audio clips in order to improve Alexa's capabilities. Recordings made by Alexa were reported possible even after the users had opted out (Day et al., 2019).
- Some companies whose security measures were far from being GDPR compliant were reported receiving a constant flow of various kinds of data sent by Alexa and other IoT devices (Hill & Mattu, 2018).

Natural Language Processing (NLP) is a branch of AI that makes possible the recognition of patterns of words and sentences in a voice and can translate them into text or act to perform the request (The Royal Society, 2017). Voice recognition systems are now more accurate, and the same technology is used to clone original voices (Lyrebird, 2017). For example, Baidu released a voice cloning system based on Neural Network, and only a few samples are sufficient to clone a voice (Arik et al., 2018).

Other examples of this potentially critical capacity are applications such as WaveNet, "a deep neural network for generating raw audio Waveforms" (Oord et al., 2016, p. 1), a powerful text to speech synthesis system (TSSS) currently being developed by DeepMind, which is able to create more realistic voices from real human speech, and Adobe VoCo, an audio editing and audio generating software from real human speech.

Figure 2-5 Intelligent virtual assistants and security and privacy risks



(Source: Chung, Iorga, Voas, & Lee, 2017)

Presented as the Photoshopping Voiceovers, Adobe VoCo can "…edit or insert a few words without the hassle of recreating the recording…(it) allows…to change words in a voiceover simply by typing new words" (Adobe Communications Team, 2016). The idea, called the "Photoshop for faking voices" (Orlowski, 2016) was less enthusiastically received by some technology researchers and the media, which did not appreciate the potential consequences of a voice manipulation application (BBC Technology, 2016) able to "…generate completely fabricated voice samples of any person they can obtain a sample from—with or without the subject's consent" (Lamphere, 2018, p. 2).

As noted in the report on the potential malicious uses of Artificial Intelligence (Brundage et al., 2018), having a system able to reproduce a voice up to the point where it cannot be distinguished from the original creates numerous issues in terms of security, ethics, and legal implications, for example in judicial cases (Gershman, 2017).

The danger of covert AI systems is also highlighted in the report *Ethics Guidelines for Trustworthy AI* (High-Level Expert Group on AI - European Commission, 2019). The document is the result of the work conducted by the High-Level Group on Artificial Intelligence set up by the European Commission. The Group of experts highlights the importance of avoiding covert AI Systems that interact with human beings and emphasises the responsibility of AI practitioners in ensuring that this is done in a transparent way (ibidem).

- Computer Vision, Image Recognition and Image Creation. Computers with ML applications are used in combination with other AI technologies to identify visual images and to link them to meaning/information already available for similar images. Tagging online images is now a common feature and similar systems are used to analyse handwritten text (The Royal Society, 2017). Image Recognition is the area which has advanced the most in the last few years, and some systems are now better than humans in identifying and categorising images, or aggregating data from different sources. Computer vision has improved immensely, leading to an industrialisation of this application (Zhang et al., 2021). However, the advancement of image recognition technologies also means that individuals are more recognisable, and their identification can be done independently of the consent of the subjects in the pictures or videos. The use of advanced image recognition technology, in combination with other technologies is a particularly powerful instrument for aggregating personal data from different sources, and this can produce some unwanted outcomes for people who, for various reasons, are not disclosing their true identities or are using multiple online identities. This is particularly concerning for political and human right activists (Bosworth et al., 2015), workers in the sex industry (Hill, 2017), and victims of revenge porn (Cole, 2017).

Two cases demonstrate the urgency of ethical regulation in the industry. A German-based Chinese programmer used Face Recognition technology to create a system allegedly able to identify 100,000 porn actresses. He crossed-referenced images from videos and from social media with the purpose of “helping” men check the trustworthiness of their partners (Fu, 2019). The misogynistic AI project had to be cancelled due to violation of European Data Protection. The programmer had to delete all data as the processing was done in the EU and it was a clear violation of the GDPR: data processing was done in one of the EU countries, collection was done without consent, and it was lacking any other legitimate basis for processing actresses’ personal data (Chen, 2019). The most recent case is Clearview, a start-up that scraped three billion images from publicly available websites and social media, created a database without the consent of the data subjects, and then sold access to the database to various organisations, such as law enforcement agencies in many countries (Rezende, 2020).

Some researchers, notably Timnit Gebru (former Google AI Ethical researcher) developed models able to predict income, pollution, crime rates and voting patterns at the local level using Google images and other data collected from other sources (Gebru, Hoffman, & Fei-Fei, 2017; Gebru, Krause, et al., 2017). She presented "a method that determines socioeconomic trends from 50 million images of street scenes, gathered in 200 American cities by Google Street View cars" (Gebru et al., 2017, p. 1). "By pulling the vehicles' makes, models and years from the images, and then linking that information with other data sources, the project was able to predict factors like pollution and voting patterns at the neighborhood level" (Lohr, 2017).

The interest in image creation has also grown. Similarly to voice creation, AI systems can generate and edit increasingly realistic synthetic images, thanks especially to the Generative adversarial networks (GANs) technology (as previously seen, an unsupervised learning type of ML). Synthetic images can be used to impersonate other people, and for deceiving or manipulating public opinion via social media (Brundage et al., 2018; Zhang et al., 2021). The use of AI to create fake news and the impact on the perception of truth was also discussed during the AI Committee meetings at the House of Lords:" [AI] risks creating a world where nothing we see or hear can be taken on trust, and where 'fake news 'becomes the default rather than the outlier" (House of Lords Select Committee on Artificial Intelligence, 2018, p. 82).

Learning from watching is another recent area of interest. For example, the learning framework called Neural Task Programming (NTP) allows learning by watching online videos (Xu et al., 2018). Another emerging area of research is based on Recurrent Neural Networks (RNN) models that, allegedly (Narayanan, 2019), predict human behaviour in crowded spaces by observing how individuals adjust their movements according to others who are present in the same space (Alahi et al., 2016, p. 3). The capacity to learn from what is perceived in videos, and the capacity to recreate actions and behaviours of the featured people and to apply them to different situations, moves the debate on DP, consent, and secondary use of biometrics data to a different level which warrants a different kind of awareness and protection, moving the focus from technical to environmental aspects.

-Searching and Organising Information: Search Algorithms and Spam Filters. Google, email spam filters, and content aggregators are the best-known examples. Searching for specific information on the Internet is probably the most widely used example of ML. Other widely recognisable examples are systems that filter spam emails, clustering and classifying content. A more recent example is a content aggregator which allows the user to search for content (such as text, chat and emails) from various users 'channels and to aggregate it into a single view. The ML learns to adapt according to the user's preferences, such as the people they communicate with more frequently and the speed of the response and opening of emails.

-Pattern Recognition and Crime Prevention. Another wide use of ML is the detection of unusual activity in data patterns, i.e., fraud and crime prevention (H.-W. Kang & Kang, 2017; O'Neil, 2016). ML technologies have shifted the approach in Financial and Crime Prevention, moving the focus from investigating past events to preventing criminal activities and making decisions about future incidents (H.-W. Kang & Kang, 2017; O'Neil, 2016). This new area has the potential to produce negative consequences for data subjects as ML can recreate patterns of discrimination contained within the datasets. If predictions are made by ML trained using biased algorithms, the decision will reinforce the original bias included in the training data (European Data Protection Supervisor, 2016), recreating patterns of discrimination. Therefore, as observed by Narayanan (2019) talking about a recidivism predicting tool analysed by Dressel & Farid (2018), ML actually predicts re-arrest as "that's what is recorded in the data. So at least some of the predictive performance of the algorithm comes from being able to predict the biases of policing" (Narayanan, 2019, p. 15). Thus, it is more a prediction tool for the praxis surrounding the data, than the data itself.

-Robotics. A robot "…the embodied form of AI…physical manifestations [that] might have sensory inputs and abilities powered by machine learning" (The Royal Society, 2017, p. 25). In relation to DP, concerns are expressed around the interaction between humans and robots, and the collection of data from the environment (from the interaction in the space and the outputs created using data from that interaction).

-Healthcare. Healthcare is one of the areas where AI is advancing more rapidly. ML has already been applied to hospital management (Callahan & Shah, 2017), it is used to support doctors in establishing diagnoses, assessing prognosis (Sayburn, 2017), and in predicting the evolution of diseases. Another area is wearable technology, which comes with specific risks, such as potential breaches, monitoring and surveillance (for example by Health Insurance companies). Wearable technology collects biometric data, special categories of data that merit higher protection as their violation can create higher risks for the rights and freedoms of individuals. While data collected by wearables is useful in having more information about patients, potential data breaches, mismanagement of data, unlawful collection, processing, surveillance, and coercion in business settings are all potential sources of concern.

DeepMind Health applied ML to medical research, specifically to diagnose eye disease (Ram, 2018), kidney failure, and breast cancer. IBM Watson used ML, in combination with Natural Language Processing (NLP), to extract information from medical notes (both written and verbal) and from research papers to provide better support to healthcare professionals.

However, the promising results coming from the application of AI in Healthcare are not without critics (Panch et al., 2019). In relation to the AI history of overpromising, and the aspirations to use AI as a substitute for human expertise, the IBM case is particularly relevant, as recently noted by Strickland (2019) who recognises the success of AI in Healthcare only in very specific situations. Positive results, mainly involving the analysis of medical images, have been attained with few AI applications. IBM Watson does not use medical images but collects and analyses written information, and it has been less successful in making sense of complex medical information (ibidem). While the case of Watson Health has highlighted the difficulties in creating an AI doctor, as the aim appeared to be more complex than originally envisaged, it has proved to be effective in providing doctors with specific information in certain areas (such as in the Genomics project). Thus, this case is emblematic of an effective and successful AI that augments the work of physicians and is less effective when AI aims at automating their jobs.

Digital symptom checker apps are other AI based healthcare systems raising concerns amongst health practitioners (Fraser, Coiera, & Wong, 2018; Oliver, 2019; Olson, 2018), especially considering the lack of resources in the NHS. The most famous is Babylon Health, which is already working with the NHS in London and providing services as a GP practice. Babylon was reported to have used software that had not been carefully checked, and to have shown exaggerated efficiency results based on non-independent assessments (McCartney, 2017). Concerns had been reported by both staff and by external doctors (Dr Murphy, 2019). Incorrect diagnoses, some of which related to cancer cases, were identified by doctors working for Babylon, who decided to perform an audit on their own (Olson, 2018). An interesting point raised by Fraser, Coiera, & Wong concerned the kinds of tests that similar systems should be subject to. Next to advocating rigorous clinical evaluation, with the publication of protocols and data sets, they encourage extending checks as to how users interact with the system "both in providing information and interpreting and acting on its outputs and…ultimately focus on real-world outcomes" (ibidem, p. 2264). This is an important point on how restricting the focus and analysis solely to data, without including the human interaction with AI systems, can lead effectively to a reduction of efficiency.

-Transport, energy, and emergency intervention. Self-driving cars can slowly become a reality, with some public regulatory bodies already issuing rules to regulate them (Somerville, 2018); drone delivery is being trialled in the UK (Amazon, 2018); and prediction made with ML can reduce traffic congestion (Asencio-Cortes et al., 2016; L. Lee, 2018). Similarly, ML can be used to analyse patterns and optimise energy consumption (Gao, 2014). It is also used for optimised interventions in emergencies. Facing an emergency requires an understanding of what "…is happening at a local level, predict what might happen next, and decide where to focus efforts accordingly" (The Royal Society, 2017). The capacity to analyse using real time data, predicting, and making immediate decisions is superior by far to other static methods of analysis (B. Kang & Choo, 2016; Zagorecki et al., 2013).

-Manufacturing and retail. Automation is impacting on production, employment (Davies & Wendes, 2014; The Economist, 2018), and personalised product recommendations.

Therefore, ML is a successful technology already used in various applications. There are some inspiring examples, such as the case of ML applied to healthcare, but also some specific challenges associated to its use.

c. ML suitable tasks

While the applications of ML are very promising, the technology is not suitable for every situation and task (Brynjolfsson & Mitchell, 2017). For example:

1. *"Learning a function that maps well-defined inputs to well-defined outputs"* (p. 1532). Typical of classifications and prediction. ML learns statistical correlations and not causal effects.

2. *"Large (digital) data sets exist or can be created containing input-output pairs"* (ibidem). The bigger and more accurate the dataset and examples, the more precise the learning.

3. *“The task provides clear feedback with clearly definable goals and metrics”* (ibidem). Clarity of goals, even if processes are less clear, performance metrics and data labelled accordingly.

4. *“No long chains of logic or reasoning that depend on diverse background knowledge or common sense” (*ibidem*).* ML is less effective in situations with complex reasoning, complex planning and multiple events. ML lacks common sense and background knowledge, or general and flexible human knowledge.

5. *“No need for detailed explanation of how the decision was made”* (ibidem). In situations when explaining the logic behind the decision is not necessary. This is a specific issue for the GPDR (see 2.3.6).

6. “*A tolerance for error and no need for provably correct or optimal solutions*” (ibidem). Prediction is a guess and contains an element of uncertainty which can be easily overlooked. The probabilistic nature of prediction can be easily misread as certain forecast, and assuming the past is similarly reproduced in the future can have some negative consequences for business strategies.

7. *“The phenomenon or function being learned should not change rapidly over time*” (ibidem). Learning with new training data should occur every time there are new changes.

8. *“No specialized dexterity, physical skills, or mobility required*” (ibidem). AI physical capabilities are still very limited.

Some of the limits are particularly relevant for organisations and will be analysed in 2.4 of this chapter (AI management).

2.2.6 Risks and data economy

The opportunities and possibilities unleashed by the adoption of AI technologies are already numerous. The number of organisations showing interest in adopting AI or already using it is growing fast. AI is also expected to produce drastic changes in various sectors and to be extremely important for the UK’s post-Brexit economy (Department for Business, Gov.UK, 2021, 2017). And yet, not only is the potential of AI in different sectors becoming clearer, but also its risks and negative consequences. Opportunities and possibilities come with some real worries in relation to job losses, socio-economic and political risks, safety risks, and erosion of privacy (High-Level Expert Group on AI - European Commission, 2019). Furthermore, new questions arise from the economic value of data, data ownership and new business models. The digital economy is booming, and personal data is an essential part. Almost everything people do leaves a digital footprint providing information on current and potential future habits and preferences. The value is in the analysis of flows of data coming from different sources, personal and non-personal data (The Economist, 2017). By using addictive systems (Eyal, 2014), often taking advantage of users' unconscious (Kreps, 2019), ad-based business models use the information to create more personalised products, boosting a market completely reliant on personal data (Pasquale author, 2015; Waters, 2018). The growth of Big Tech, particularly the FANG group (Facebook, Amazon, Netflix, Google) (The Economist, 2018a), seems unstoppable, making them key players in future technocracies (Helbing et al., 2017).

Discussions on AI and biases have increased significantly. Concerns about privacy have grown (Christl, 2017; Pasquale, 2015; Waters, 2018; Zuboff, 2019) especially after the scandal caused by Cambridge Analytica, which also involved Facebook (Greenfield, 2018). The case was emblematic. By using behavioural communication, psychographic and data analytics (Concordia, 2016), the case was an example of new business models based on data. Demographic (i.e., age, ethnicity), psychographic (i.e., consumer data) and personality behaviour (i.e., openness, fear) data was aggregated and used to profile an audience and send "persuasive" messages for political communication. Their business model was not unique. The specific role of ML in "creating information from data" (The Royal Society, 2017, p. 90) is under scrutiny, and how data is used by organisations needs to be better understood. Traditional approaches to DP do not always appear sufficient to protect the rights of data subjects when their data is considered with other information. For instance, ML has the capacity to link and aggregate personal data from different sources (potentially identifying individuals), facilitate the re-identification of anonymised data, and the identification of special categories of data (e.g., sexual orientation), warranting higher protection due to the potential for discriminatory use.

Therefore, the risks are multiple, the adoption is increasing, but so are the debates and the regulations. Various countries, particularly the US, China, the EU, and the UK, have increased AI adoption or are planning to use it as key technology in their economic strategic plans (Allen & Chan, 2017; Center for AI and Digital Policy, 2020; CIFAR Canadian Institute for Advanced Research, 2017; Hall & Pesenti, 2017; Zhang et al., 2021).

Technologists, politicians, economists, and philosophers are debating the implications of AI for societies (Tegmark, 2017; Department, Business, & Strategy, 2017; Bostrom, 2014) and international organisations, such as the European Union and the United Nations, have published regulations and reports on various aspects of AI technologies, (Delvaux, 2017; ITU, 2017; European Commission, 2021; High-Level Expert Group on Artificial Intelligence, 2019).

The protection of personal data has massively evolved in recent years, and the increased awareness of the risks of AI has contributed to the inclusion of requirements into the legislation. This research focuses on the European tradition, and it looks at how DP has evolved in the continental and UK legislation. This will be presented in part 2.3.

2.2.7 Gaps in the literature addressing AI

The above literature review has presented the exponential growth of AI. The literature is mainly focused on the evolution of the technology, the growth of data, and the increasing range of AI products. Some gaps can be identified. Research into how the technology is being used is still scarce. There is a common lack of consideration for the conditions under which AI is adopted or for the people choosing and using AI for their organisational needs. There is insufficient discussion of the wider social and organisational context surrounding the use of this technology, and where, how, why and by whom it is used. This research aims to fill these gaps by shifting the focus to the use, context, and the people who are managing AI/ML. This is done by extending the scope of empirical research, going beyond mere technological concerns, to explore the organisational context, including a focus on people who are shaping its use, their understandings, expectations, and perceptions. In doing so, all the factors impacting on usage, such as organisational dynamics, risks, and constraints influencing management's decisions are also taken into consideration.

2.3 Data protection (DP)

2.3.1 The evolution of the right to DP in the European legal order

This part presents the evolution of the concept of DP. It starts with a discussion of the right to DP in Conventions published by the Council of Europe, and it then illustrates the evolution of DP right in the European tradition. It presents the previous DP law, the current regime with the GDPR, the UK Data Protection Act 2018 and the UK GDPR. It then ends with a presentation of the current debate on fairness, accountability and transparency and their links to the GDPR.

2.3.2 Council of Europe. The European Convention of Human Rights (ECHR)

The right to DP, intended as the right to the protection of personal data, and the right to privacy – the right to respect private life – are closely related but they are two distinct rights. The United Nations General Assembly first codified the right to privacy in the Universal Declaration of Human Rights/UDHR (United Nation General Assembly, 1948). The Declaration stated the right not to be subjected "to arbitrary interference with his privacy, family, home or correspondence…" (Art. 12). The Council of Europe (Figure 2-6) adopted the European Convention of Human Rights/ECHR (The European Convention on Human Rights (ECHR), 1952) after a few years, including the right of privacy with a similar text in Art. 8.

With the development of technology, a new concept around the collection and use of personal data started to emerge in the legislation and jurisdiction of some European countries. National laws have been adopted since the '70s in Sweden, Germany, the Netherlands, and the UK, mainly aimed at controlling the processing done by public institutions and large organisations (European Union Agency for Fundamental Rights, 2018).

In Germany, the right to informational self-regulation, intended as "the authority of the individual to decide himself…when and within what limits information about his private life should be communicated to others" (Rouvroy & Poullet, 2009, p. 45) was identified by the German Federal Constitutional Court in 1984. The pronouncement of the court annulled the decision to conduct a general population census and identified a new basic right which was "…the legal anchor for Data Protection in the German constitution… the most important decision in the history of German Data Protection" (Hornung & Schnabel, 2009, p. 84).

Figure 2-6 Map of the Council of Europe 47 member states

(Source: Council of Europe)

According to Hornung and Schnabel, the right is linked to the concept of personality right, the idea that individuals have the right to "develop a free and self-determined personality" (ibidem, p. 86), and for this reason the right to informational self-regulation can be applied only to individuals and not to legal entities.

2.3.3 Council of Europe. Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data (Convention 108)

The Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data/Convention 108 (European Council, 1981) adopted by The Council of Europe was the first legislation on DP approved at European level. It contained provisions on data processing carried out by private and public organisations, inclusive of law enforcement authorities. Some of the provisions will later be expanded in the GDPR, such as the definition of personal data, special categories of data, fairness, and lawfulness, right to information, data security, transborder flows of personal data. The convention was amended in 2018 to enhance the protection of the rights and freedoms of data subjects from the danger of new technologies. New provisions included those introduced by the GDPR, e.g., biometrics (Amendments to Convention 108, 2018).

Adopted by all members of the European Council (47 members), plus another four non-members, Convention 108 has the "…potential as a universal standard, together with its open character, to serve as a basis for promoting Data Protection at global level" (European Union Agency for Fundamental Rights, 2018, p. 26). Thus, the convention is the only international agreement on DP and has the potential to expand the right to other countries.

2.3.4 European Union: Data Protection Directive 95/46/EC

The Data Protection Directive 95/46/EC was the first legislation approved by EU institutions. Essential for the free flow of data within the internal market, the directive provided a high level of DP (Publication Office of the European Union, 2018) facilitating the free movement of goods, capital, services, and people (Four Freedoms) between Member States. As EU Directives are EU secondary laws (creating obligations for Member States but not for individuals), Member States must transpose them into internal laws. The Data Protection Act (DPA) 1998 (Gov.UK, 1998) was the bill that transposed the Directive into internal law.

2.3.5 European Union: Charter of Fundamental Rights of the European Union

The European Communities (EC) treaties and the Maastricht treaty that created the European Union did not make any provisions for Human Rights. The Charter of Fundamental Rights of the European Union (European Parliament and Office for Official Publications of the European Communities, 2000) was the first document on Human Rights of the European Union. It was adopted via a primary EU legislation which, as with treaties, is approved voluntarily and democratically by all EU member countries (European Union, 2019).

The Charter protects the Human Rights and fundamental freedoms of EU Citizens and Residents in the EU: Dignity, Freedoms, Equality, Solidarity, Citizens' rights, Justice. The document was comprehensive of principles derived from various sources, including rights from The European Convention on Human Rights (ECHR), cases from The European Court of Justice, and Principles of Jurisprudence of Member States.

The Charter was not legally binding and came into force with the Lisbon Treaty (European Union, 2007) which specifically referenced the Charter in Art. 6, creating an obligation for Member States to conform to those principles.

The Charter covered the right of privacy “Respect for private and family life: Everyone has the right to respect for his or her private and family life, home and communications.” (Art. 7), and it recognised DP as a new right, closely linked to the right of privacy but independent from it:

1. Everyone has the right to the protection of personal data concerning him or her.

> 2. Such data must be processed fairly for specified purposes and on the basis of the consent of the person concerned or some other legitimate basis laid down by law. Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified.
>
> 3. Compliance with these rules shall be subject to control by an independent authority (Art. 8).

Linking DP to the right of privacy but, de facto, separating it, is quite unusual among international Human Rights documents (Lynskey, 2014), where DP is usually derived from the right of privacy. Lynskey recognises the importance of the distinction, as "[t]he Right to Data Protection provides individuals with more control over personal data than the Right to Privacy" (ibidem, p. 29). However, she sees this combination as a cause of confusion and ambiguity, for example, some commentators saw the inclusion in the Charter as the recognition of a right that was, de facto, already conveyed in the Data Protection Directive, while others considered it as an extension of the application of the Directive to areas previously excluded from it.

A lack of consistency in EU law is also noted by Erdos (2015). For example, the jurisprudence of the Court of Justice of the European Union (CJEU) "has often elided this new right with that of the more traditional right to privacy" (ibidem, p. 374). However, Erdos reported a specific court case as the key moment when DP was seen as having a different status from the right of privacy. The case Google Spain SL, Google Inc. v. AEPD, Mario Costeja González [2014] ECR 1-62012 (Case, 2014) was the pivotal case which recognised the right of individuals to request search engines to delete personal information (in specific circumstances), creating the basis for the future right to erasure or right to be forgotten (see 2.3.6).

Although linking the two rights can lead to some unintended ambiguity, listing DP as a distinct right created a legal obligation for Member States and a legal protection for individuals in the EU.

2.3.6 The European General Data Protection Regulation (GDPR)

The European General Data Protection Regulation/GDPR (Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such

Data, 2016), became enforceable in the EU on 25 May 2018 (Art. 99 GDPR) and in the EFTA area (Iceland, Liechtenstein, Norway, and Switzerland) (EFTA, 2018) after four years of preparation. The text, a compromise between Parliament and the Council and Commission of the European Union, repealed the Data Protection Directive 95, creating a more uniform and directly applicable DP regime in the EU as regulations are automatically enforceable after two years.

The GDPR was necessary in order to modernise the legislation to protect the rights and freedoms of individuals in the context of the digital economy. By strengthening the right to DP as an independent right, it introduces a radical change of perspective shifting the focus from the single market to human rights (Dresner, 2014).

The GDPR develops core principles from the Directive, from national laws and from the Council of Europe's Convention 108 (European Union Agency for Fundamental Rights, 2018). It shifts the focus onto organisations, introducing the principle of accountability and new strict requirements, and provides organisations with a more flexible mechanism for operating their business in different EU countries (One-stop-shop art. 56.1).

A) Principles

The GDPR outlines seven principles that must be at the core of any personal data processing (Art. 5)

1. Lawfulness, fairness, and transparency

- Lawfulness (Art. 6). Data can only be processed lawfully if there is at least one of the following elements: consent of the data subject; necessary for a contract or to enter into a contract; compliance with a legal obligation; necessary to protect the vital interest of the data subject or another person; necessary for the legitimate interest of the controller or a third party, when they are not overridden by "the interests or fundamental rights and freedoms of the data subject which require protection of personal data…" (Art. 6.1(f)).

- Fairness. Data must be processed in a fair and transparent manner (Art 5.1(a)). Controllers "must be able to demonstrate the compliance of processing operations…[they] must not be performed in secret …and data subjects should be aware of potential risks… (European Union Agency for Fundamental Rights, 2018, p. 118). In interpreting the principle of fairness, the Agency make a significant observation for this research: "…the principle of fairness goes beyond transparency obligations and could also be linked to processing personal data in an ethical manner" (ibidem, p. 119). The example used by the Agency is especially important for this research. A university research department realises that data gathered for a scientific project could be also used for another project carried out by another team within the same department. As the controller is the same (university), and the purposes of the two projects are compatible, the controller could process it lawfully according to Art. 5.1b. However, "…the university informed the subjects and asked for new consent, following its research ethics code and the principle of fair processing" (ibidem). Therefore, a clear distinction is made between lawfulness and fairness, with fairness being assessed separately. The example is significant for AI. As highlighted by one of the participants (Part. 6), many organisations are acquiring the technology for a specific aim, but often other uses are identified after the implementation, with or without the original data. This can have massive implication for DP and fairness.

- Transparency. The general obligation on the part of organisations to inform data subjects on how data will be used.

2. Purpose limitation. Processing is done for "specific, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes…" (Art. 5.1b). Further processing is possible in specific cases (archiving purposes in the public interest, scientific or historical research purposes or statistical purposes) and only if the new purpose is compatible with the original one ('purpose limitation'). Compatibility should be assessed considering the links between the two purposes (Art. 6.4), the context of collection, the expectations of data subjects, nature of data, potential effect on data subjects, appropriate safeguards (such as encryption and pseudonymisation) (European Union Agency for Fundamental Rights, 2018).

3. Data minimisation. Processing of only data that is "adequate, relevant and not excessive" in relation to that purpose (Recital 50).

4. Data accuracy. Controllers must verify accuracy of data, also via regular checks, and rectify and erase without delay (Art. 5.1(d)). Considering the importance of data quality for ML, satisfying this principle is even more important in the case of AI.

5. Storage limitation. Data must be deleted or anonymised as soon as the purpose is served (Art. 5.1(e)).

6. Integrity and confidentiality (security). Organisations must take appropriate security measures (technical or organisational) to prevent negative consequences for data subjects such as protection against unauthorised access, loss, damage, or destruction (Art. 5.1(f)).

7. Accountability. The GDPR new principle. Organisations are responsible and have to demonstrate compliance with all the above requirements. Accountability can be a challenge for organisations using deep learning and black box (more in 2.4.3).

The principles of fairness, accountability and transparency and their relation to ML are better explored in 2.4.1.

B) GDPR key elements

The following sections give details on the key elements of the GDPR, with some implications for ML.

a. Territorial scope and international reach (Art. 3)

The Regulation applies to organisations who are processing personal data of individuals in the EU, and specifically when at least one amongst data subject, controller, and processor is in the EU:

1. Data Subject. The applicability is not dependent on citizenship or formal residence, and it applies to all individuals who are in the EU, permanently or temporarily (i.e., on holiday or flying across).

2. Organisations (controllers/processors) whose activities or "effective and real exercise of activity through stable arrangements" (Recital 22) are based in the EU and process data of people (in or outside of the EU). The GDPR applies irrespective of processing done in or outside the EU.

3. Organisations (controllers and processors) based outside the EU that process data of individuals who are in the EU. This applies to entities who are offering goods or services (irrespective of a required payment) to individuals in the EU, or who are "monitoring…their behaviour, as far as their behaviour takes place within the Union" (Art. 3.2b). This specific provision is particularly relevant for Non-EU Organisations who employ AI/ML systems to monitor and profile individuals who are in the EU. The case of Internet profiling is specifically mentioned as an example of monitoring, when: "natural

persons are tracked on the internet including potential subsequent use of personal data processing techniques which consist of profiling…in order to take decisions…or for analysing or predicting…preferences, behaviours and attitudes." (Rec. 24). Furthermore, the GDPR prescribes that controllers or processors not established in the Union must designate a representative in the EU (Art. 27).

b. Definition of personal data

The definition of personal data is expanded to include any information that can identify a person, such as "a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person (Art. 4.1). Digital identifications (fundamental to location-based marketing) are also considered personal data, such as "online identifiers provided by their devices, applications, tools and protocols, such as internet protocol addresses, cookie identifiers or other identifiers such as radio frequency identification tags" (Rec. 30).

Even though the definition of personal data has been extended, some areas are excluded. This can be problematic considering some new AI applications:

- Anonymous information or personal data rendered anonymous are excluded by the GDPR (Rec. 26). As the identification can be done directly and indirectly, the Regulation is very careful with regards to future technological developments and potential re-identification tools, and it clarifies that it should be taken "into consideration the available technology at the time of the processing and technological developments" (ibidem).

Particularly interesting with regards to the use of algorithms in data anonymisation, is the analysis made by Elliot et al. (2018), which connects the anonymization in Rec. 26 to data sharing. They see anonymization not only as a privacy tool, but also as a "procedure that adds to the business case for data sharing (p. 3), and that should be done as "functional anonymization", a compelling new practice that reduces the risk of re-identification. They link disclosure risk to data environment and argue that there are elements in the environment that can be controlled by the controller, such as "skills and…motivation of people, structure, processes and the infrastructure in which the data resides" (p. 14). As anonymization of data can be insufficient, they suggest a holistic methodology. "[A] combination of technological and nontechnological methods applied in an integrated fashion, is likely to be more effective than an approach that casts anonymization as exclusively technological, with nontechnological processes bolted on as an afterthought." A fascinating new approach to anonymization which is more aware of human-technology interactions and contexts.

- Afterlife data rights. Personal data of deceased people are also excluded by the application of the GDPR (Rec. 27), leaving it to the discretion of Member States to create a minimum protection (Harbinja, 2017). Considering the evolution of AI, this is problematic (Graziano, 2016). There are projects for creating systems using personal data (such as voices) to recreate " convincing digital surrogates of the dead" (Matei, 2018), or "mental duplicates (chatbots crafted from personal data" (ibidem), a "digital identity of deceased persons" called "Augmented Identity" (Sofka et al., 2017). "Algorithms would use recorded data to answer questions posed from beyond the grave", (ibidem, p. 178) creating a "new form of inter-generational collective intelligence" (Tynan, 2016) or a digital incarnation of the dead as chatbots (Collins, 2021).

-Consciousness. Research is also being conducted on the possibility of replicating consciousness. As noted by Graziano (2016), if we think of our brains being scanned to recreate a virtual post-mortem replica for a future biometric self, does that copy constitute personal data? And if it is modified by ML systems, is it still us or not? New questions are arising and are also related to the future of personal data. This is not a debate any longer confined to the realm of philosophy, but for example it is one of the main goals of The Human Brain project (European Institute for Theoretical Neuroscience (EITN), 2018). Could consciousness be considered personal data? Would that be regarded as biometric data? Will the consciousness of a dead individual be included in future DP legislation? The GDPR provision is not sufficient to regulate afterlife data rights and the progress in ML requires a different approach.

- Special Categories of Personal Data. Some categories of personal data deserve special protection, as they "could create significant risks to the fundamental rights and freedoms" (GDPR, Rec. 51). Their processing is prohibited and only allowed in specific cases, such as: explicit consent of data subject (controllers cannot use legitimate interest as a legitimate basis for processing); specific provision by the Member States in employment and social security law; vital interest of an incapable data subject; legitimate activities of special entities; data made public by the data subject; legal claims and court cases; public interest; preventive and occupational medicine (Art. 9.2).

The list of categories (Sensitive Data) included in the Directive 95 (information related to racial and ethnic origin, political opinions, religious beliefs or other beliefs of a similar nature, trade union membership, physical or mental health or condition, sexual life) is expanded with new categories: sexual orientation, and "…genetic data and biometric data for the purpose of uniquely identifying a natural person…" (Art. 9.1).

The inclusion of genetic and biometric data is particularly relevant for ML systems. For example, Recital 51 prescribes that the processing of photographs should be considered processing of special categories of personal data only when processed through a specific technical means allowing the unique identification or authentication of a natural person. Face recognition systems are therefore included in this provision, prompting Facebook to reintroduce its Face Recognition system in the EU (Kuchler, 2018).

- Emotion AI or Emotion Detection and Recognition (EDR). A growing area in ML is Emotion AI (Markets and Markets, 2020; P. Nilsson, 2018), based on new "perceptual" systems that are claimed to identify and measure emotions in individuals. Various elements are observed, and data gathered in order to make a decision, for example faces, heartbeat, body language, voice, speech pattern.

Applications are being developed by various companies for different sectors, e.g., retail (Guha et al., 2021), recruitment (Su et al., 2021), education (Nimala & Jebakumar, 2021). Different tools are used. For instance, combinations of Magnetic Resonance Imaging (MRI), Electroencephalogram (EEG) and eye-tracking methods (Trueimpact, 2017), brainwaves, facial analysis, and sound analysis (Unruly, 2018), or eye tracking to measure emotions (Real Eyes, 2018; Sticky, 2017).

The most important company is Affectiva, a multi-modal emotion AI (Mcduff et al., 2013) using a combination of computer vision, speech science and DL: "The API analyzes not what is said, but how it is said, observing changes in speech paralinguistics, tone, loudness, tempo, and voice quality to distinguish speech events, emotions, and gender" (Affectiva, 2018). Critical was Silent Talker/iBorderCtrl, an AI based system measuring micro-expressions or "biomarkers of deceit" ((Sánchez-Monedero & Dencik, 2020, p. 1) used to assess the mental state of travellers. It " empower[ed] border agents to increase the accuracy and efficiency of border checks" (EU Directorate-General for Research and Innovation, 2018). This contested Augmented AI system was subjected to a transparency lawsuit seeking the release of, amongst others, the ethical evaluation of the project (Lomas, 2021).

Emotions are not clearly included amongst Biometrics, and the application of AI to Emotions and feelings can lead to powerful consequences. Are emotions mental health data? It is not clear, and the legislation gap is a cause of concern. If emotions were to be included within biometric data (as part of mental health data), then the GDPR should be applied to the new data that is gathered and processed via Emotion AI or Empathic AI. “We know how you feel” was the title of an article published on the work done by Affectiva (Khatchadourian, 2015), which presented some of the implications in reading another person’s feelings from facial expressions. The scrutiny of the relationship between emotions as personal data and AI systems is an urgent need.

c. Consent

Consent is one of the lawful conditions for processing personal data (Art. 6) amongst: contractual necessity, legitimate interest, public interest, vital interest, legal obligations. Consent is the most debated requirement with regards to AI. The GDPR defines consent as: “any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action…” (Art. 4.11). Conditions for a lawful consent are clearly indicated by Art. 7, and later clarified by some guidance published by the EU Working Group 29/WG29 (Article 29 Data Protection Working Party WP 223, 2018), and the ICO (ICO, 2017, 2018), for example:

- Free/Freely given implies real choice and control given to the individual (Article 29 Data Protection Working Party, 2018, p. 5). Consent should not be an obligation but a real option without negative consequences. Consent is not valid if: requested in return for a service; negative consequences are derived from denying it; there is power imbalance between the parties (i.e., public authority or employment).

- Specific. Consent is given for a specific purpose, it can be withdrawn at any time, and it shall be as easy to withdraw as to give consent (Art. 7). Processing and retention is possible until it satisfies the original purpose and cannot be used for incompatible purposes (Art. 5.1b). This is a safeguard against the “Function Creep” phenomenon, when data is used for other purposes (Article 29 Data Protection Working Party WP 223, 2018), a risk that, as recently seen in the Cambridge Analytica case, can be highly damaging for individuals, for business continuity, and for democracy. 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 (‘purpose limitation’) (Art. 5.1(b).

The possibility to re-process data for a “compatible purpose”, and for archiving purposes, is crucial for AI and ML technologies. This inclusion was seen by some authors (Mayer-Schonberger & Padova, 2015) as the result of political lobbying done by Big Tech companies in Brussels who asked for consent used for a wide purpose, for keeping data for statistical purposes, or for re-using data after anonymisation.

- Informed. Transparent information to be provided to data subject about identity controllers, purpose, type of data and how it will be used, automated processing, and data transfer.

- Unambiguous indication of consent (Rec. 32 and Art. 4.11). Statement can be written or recorded (oral statement). No pre-ticked boxes or silence or inactivity can be considered valid manifestations of consent, and it must be recorded and stored for audit purposes.

- Explicit consent for risky situations. For example, in case of special categories of data, profiling and automated decision making (Art. 22), consent must be given via a clear affirmative action…” (Art. 4.11) (i.e., written statement, ideally signed by data subjects).

d. Controller and processor

The GDPR is very prescriptive. Controller defines the purpose and means of processing (Art 4. (7)), and processor processes personal data “on behalf” of the controller (Art. 4 8). Processing is regulated by a written contract (28.3). Processors must be able to guarantee high technical and organisational measures (Art. 32.1) (such as testing, confidentiality, security, integrity, availability and resilience of processing systems and services, pseudonymisation and encryption, ability to restore the availability and access to personal data in the case of physical or technical incident and not sub-processing without authorisation). Many requirements are easier for big companies, as they have more resources to invest in organisational and technical changes (Webber, 2016).

The difference between controller and processor becomes particularly interesting with AI technologies. Vendors are bringing in and supporting organisations with implementation. Their role as mere processors is disputed, as also highlighted by the report conducted on behalf of the Dutch Ministry of Justice and Security (Roosendaal, 2018).

The relationship between controller and processor/organisation and vendor is a key feature in AI management; this will be presented in 2.5.

e. Data subject rights

The rights of Data Subjects are several, such as: right to access personal data (Subject Access Request/SAR, Art. 15), right to data portability, and right to have data exported onto a machine-readable format and transferred to another controller (Art. 20), right to rectify inaccurate personal data (Art. 16), right to erase personal data (right to "be forgotten", Art. 17), right to restrict personal data in certain circumstances (Art. 18) and right to object to processing (Art. 21).

There are some issues in relation to the processing done with AI. The right to erasure can be requested for data provided by individuals, but not for data created from observing behaviour offline and online, as done by Facebook or LinkedIn, who "…make heavy use of observed behaviour…" (Edwards & Veale, 2017, p. 69). Metadata such as likes, geolocation, or clicked links, or inferences that could be used to identify a data subject are not included in this right. However, the right to portability can be requested for personal data provided by the data subject and for metadata, but not for inferences, "that are then drawn from that data by the Machine Learning or profiling system itself" (ibidem, p. 74).

f. Data Protection Officer (DPO)

DP experts who inform, advise and monitor GDPR compliance. They are required for some organisations, for example, public authorities (except courts acting in their judicial capacity); for organisations who carry out large scale or regular and systematic monitoring of the behaviour of individuals, or large-scale processing that could lead to high risk, as in, for example, of special categories of data or data relating to criminal convictions and offences. Organisations involved in highly risky processing are, for example, big data analytics. An indication of large-scale processing is offered by Recital 91, which refers to a "considerable amount of personal data at regional, national or supranational level and which could affect a large number of data subjects, and which are likely to result in a high risk".

g. Requirement for breaches, reporting and sanction regimes

-Data breaches (i.e., cyberattacks or unauthorised access) must be reported within 72 hours of the organisation becoming aware of it, both to Regulators and to Individuals, "unless the personal data breach is unlikely to result in a risk to the rights and freedoms of natural persons" (Art. 33), or if "the data is anonymised or encrypted". Organisations not reporting a breach will incur a double fine (for the actual breach and for the lack of reporting).

-Sanctions. Organisations are required to demonstrate how they are complying with the GDPR, and DP authorities can assess how they are using personal data (via audit and inspections) and impose sanctions.

- Fines of up to €10m or up to 2% of the total worldwide annual turnover in the case of minor breaches (Art. 83.4), such as child consent, lack of measures for DP by design and by default, or missing contract between controller and processor (Art. 83.4).
- Fines of up to €20m or up to 4% of the total worldwide annual turnover in the case of:

  -major breaches, such as: unlawful process (e.g., lack of consent or conditional consent, lawful basis for processing, or different purpose), lack of security (e.g., no staff training or inadequate cyber security), or breaches relating to Special Categories of personal data (Art. 83.5).

  -non-compliance with an order by the supervisory authority (Art. 83.6).

Big data companies are particularly at risk, according to the ICO, as "…the high level of replication in big data storage and the frequency of outsourcing the analytics increase the risk of breaches, data leakages and degradation…(and)…links between the different datasets could increase the impact of breaches and leakages" (ICO, 2017, p. 49). Similar considerations can be made about IoT and AI technologies.

h. Data protection impact assessments (DPIAs), privacy by design (PbD) and privacy by default

Data protection impact assessment (DPIAs) (Art. 35) is a tool used to evaluate potential consequences of processing in high-risk processing (i.e., using new technologies). It must be performed in the case of:

- Systematic and extensive processing, for example profiling (Art. 35.3. (a)).
- Large scale processing of special categories of data (Art. 35.3. (b)), or personal data in relation to criminal convictions or offences, including "processing a considerable amount of personal data at regional, national or supranational level…that affects a large number of individuals; and involves a high risk to rights and freedoms" (Recital 91).

- Large scale, systematic monitoring of public areas, using for example CCTV (The ICO, 2017).

As ML is likely to be highly risky for individuals, a DPIA is needed for ML systems.

Article 29 Data Protection Working Party published guidelines on DPIA (Article 29 Data Protection Working Party, 2017a) on how to assess the high risk. The document lists nine criteria to be used to evaluate the risk: (1) Evaluation or scoring; (2) Automated decision-making with legal or similar significant effect; (3) Systematic monitoring; (4) Sensitive data; (5) Data processed on a large scale; (6) Datasets that have been matched or combined; (7) Data concerning vulnerable data subjects; (8) Innovative use or applying technological or organisational solutions; (9) Data transfer across borders outside the European Union; (10) When the processing in itself "prevents Data subjects from exercising a right or using a service or a contract" (ibidem, p. 8-9). If less than two criteria are met, the DPIA might not be required.

Edwards & Veale (2017) make a detailed analysis of the implications of DPIAs for ML systems, anticipating the need to perform a DPIA for all ML systems, as also indicated by the ICO (2017) and by Article 29 Data Protection Working Party. Edwards notes that, as per GDPR requirements, almost every new ML system will be considered as high risk, and therefore needing a DPIA, and as DPIAs and PbD are less common in the private sector, it is important to understand how "these can be folded into commercial development time cycles and profit motivations, and not just become tick-box bureaucracy" (The UK Parliament, 2017).

Also of relevance is a report on discrimination and automated decision-making written by Borgesius (2018), who observes that organisations have to consider unfair discrimination in their DPIA (as per GDPR). However, as non-specific impact assessments for AI currently exist (Reisman et al., 2018a), Borgesius hopes that "Equality Bodies and human rights monitoring bodies could help to develop a specific method for a human rights and AI impact assessment" (Borgesius, 2018, p. 31), by involving different people and different disciplines.

The Brussels Laboratory for Data Protection & Privacy Impact Assessments does not consider the GDPR framework strong enough. They identify some weak points (Kloza et al., 2017), such as: lack of indication about process and methodology; a certain amount of discretion in choosing when and how to perform DPIA: "…by the very nature of the risk management process, data controllers choose, inter alia, the method of assessment and measures for risk mitigation" (ibidem, p. 3); lack of clarity about the advice the DPO should provide organisation, as per Art. 35.2. They suggest broadening the scope, by including privacy impact assessment for intrusive cases outside of the scope of GDPR, formulating different methods for performing them (respecting local differences), and an active role of the EDPB and national Regulators in sharing info. Noticeable in the report is the complete absence of any reference to AI.

Privacy by design and privacy by default are two "privacy-enhancing technologies" (PETs), made obligatory by the GDPR, which aim at reducing the amount of personal data collected by organisations. They are proactive tools that prevent (Art. 25.1) or reduce (Art. 25.2) the amount of data, reducing risks and accountability. By default, the highest privacy setting is automatically applied to a new product, and by default, personal data should be kept only for the time necessary.

The correlation between privacy by Design and Data Subject rights is interesting with regards to ML. The paper by Veale, Binns, and Ausloos (2018) is particularly compelling, and some of their findings are thought-provoking especially with regards to AI/ML. For example, the possibility that a competitor more technically advanced than a controller could re-identify data previously being treated via privacy design strategies, or that data subjects would not get access to their data via subject requests.

i. Profiling and automated decisions

As profiling and automated processes can significantly reduce the rights and freedoms of individuals, the Regulation introduces the definition of profiling, new rights for data subjects and new obligations for controllers (rights of explanation/information and right to request human intervention).

-Profiling is "…any form of automated processing of personal data…to evaluate certain personal aspects…in particular to analyse or predict…performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements" (Art. 4 (4)). For example, collecting data for finding correlations, and classifying people in order to create a virtual picture of what they are, how they behave and what they are like, no matter the purpose of the processing (e.g., credit checks, websites analytics or ad micro-targeting).

-The right to refuse a decision made only via automated processing. Individuals cannot refuse to be subjected to the processing, but they can refuse, in general, the decision made via automated processing.

-The right to request human intervention in the case of decisions which produce legal effects, or which similarly significantly affect the individual (Art. 22.1), such as citizenship, benefits, or online credit application or e-recruiting practices…profiling…that analyse or predict aspects concerning the Data Subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her (Rec. 71). Examples of “legal” or “similarly significant effects” have been clarified by guidance published by Working Group 29, e.g., an effect on legal status or on the rights deriving from a contract, cancellation of a contract or refusal of welfare benefits or citizenship (Article 29 Data Protection Working Party, 2018). Similarly significant is the reference to decisions lacking legal effects but which can affect individuals a great deal, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention (Rec. 71). Some exceptions exist, as human intervention cannot be requested when the decision is:

- Necessary for entering or performing a contract between the parties.
- Authorised by the European Union or by the Member State (for example in case of tax evasion prevention).
- Based on the explicit consent of the data subject to the automated processing (Art. 22. 2).

Therefore, the GDPR does not restrict profiling in general, but only the decisions made via automated processing, and it does not restrict all automated decisions in general, but only those producing legal effects or similarly significant effects.

- Right of explanation about the processing of their data, to obtain an explanation of the decision and to challenge it (Rec. 71). The right to “express their point of view” is also given to data subjects in the case of processing personal data for the performance of a contract, and processing following an explicit consent given by the data subject (Art. 22.3).

The right to request human intervention and the right of explanation are fundamental in providing more safeguards to data subjects. Considering the issue of opaque algorithms, typical of DL, and the expansion of the algorithm-driven economy, both are valid instruments for increasing control, transparency, and accountability.

The existence of a new right to information/explanation of decisions made by AI algorithms has been the topic of the “algorithmic war stories” (Edwards & Veale, p. 64), a lively debate amongst researchers and practitioners (Goodman & Flaxman, 2016; Edwards & Veale, 2017; Selbst & Powles, 2017; Casey, Farhangi, & Vogl, 2018) which shows how the concept of transparency is intended in the Regulation. The details are presented in Section 2.4.4, with some specific considerations on AI/ML systems.

The GDPR is a fundamental legislation in the history of DP. It has influenced new DP legislation beyond the EU, such as the California Consumer Privacy Act of 2018 CCPA (2018), the LGPD in Brazil (Raul, 2018), and the Washington Privacy Bill (Cesaratto, 2019).

2.3.7 Data Protection Act 2018

Data Protection Act 2018 (c.12) (Data Protection Act 2018/DPA) replaced the Data Protection Act 1998, defining the UK DP framework. It includes:

1. GDPR Specifications and derogations. It provides specifications and details left by the GDPR to States to define. For example: the notion of “Public Authority” (ibidem, Section 7); the minimum age for a child to consent (13 years, Section 9); special categories of personal data and criminal convictions (Section 10); situations where individuals’ rights are limited (e.g., research, historical or statistical purposes; health, social work, and education (Section 15); accreditation of certification providers by the ICO, and conditions for issuing a certification (Section 17); safeguards in archiving, research, and statistical purposes, e.g., if potential substantial damage or distress, processing data is not permitted (Section 19).

   The Act extends the GDPR standards to areas originally excluded by the Regulations, for example immigration control (Data Protection Act 2018, p. 13).

2. Law enforcement processing. The GDPR provisions do not apply to personal data processed by criminal justice agencies (such as the Police, criminal courts, prisons, or the intelligence services), as another piece of EU legislation was specifically created for these cases. The Directive 2016/680 or Law Enforcement Directive (LED) (Parliament et al., 2003) regulates processing of personal data by authorities for “the purposes of the

prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security" (ibidem, Art. 1). As previously seen, the EU directives require internal legislation to transpose them inside the territory of the state. The DPA 2018 implements the Directive 2016/680 in the UK and defines rules in relation to lawful bases, consent, privacy notices, individuals' rights, breaches, DPOs etc.

3. Information Commissioner's Office/ICO. It defines roles, obligations, and enforcement regulations, such as safeguards for issuing fines, rules on the re-identification of de-identified data offences, and enforcement.

4. National Security and Intelligence Services. The Act harmonises the UK legislation with the Data Protection Convention 108 by the Council of Europe (Council of Europe, 1981).

The Act also provides the safeguards for automated decision-making, clarifying the meaning of "based solely on automated processing" (GDPR, Art. 22.1), as lacking "meaningful" human involvement in the decision-making process. As seen in Section 2.3.6, data subjects have the right to object to a decision made only via automated decisions (including profiling) when this has legal and significant effects (GDPR, Art. 22). This right is not contemplated in the case of a contract, when authorised by the law or based on explicit consent (Art. 22.2).

According to the specific case for processing, the GDPR establishes different kinds of safeguards. When processing is based on a contractual necessity or an explicit consent, controllers must implement suitable measures to safeguard the data subject's rights, freedoms and legitimate interests (GDPR, Art. 22.3), e.g., the right to human intervention, and to express their point of view or contest the decision (ibidem). When processing is required or authorised by a law, there is no need to specify the need for automated processing in the text of the law (GDPR 22 (2) (b)).

The DPA 2018 clarifies that the controller must notify the data subject that a decision has been taken based solely on automated processing (Data Protection Act 2018, p. 9). Data subjects can respond within 30 days requesting the controller "reconsider" or take a new decision involving humans, then controllers have 30 days to consider the request, comply, and inform the data subject in writing (ibidem, p. 10).

Another important new aspect is the obligation to keep a data log (Section 62) of automated processing. Logging is required when data is collected, altered, consulted, disclosed, transferred, combined, and erased. Logging can be an important instrument that controllers can use in auditing processors and checking lawfulness in their processing. However, this is not highlighted in the Act. The importance given by the GDPR to controllers' power on processors seems to have faded in the Act.

Therefore, the DPA 2018 is a fundamental document. Particularly important for AI are the specifications on automated processing, specifically the new terms for organisations to reconsider decisions and/or involve humans, and the obligation to keep a log.

2.3.8 UK GDPR

The UK GDPR (UK Department for Digital, Culture, Media & Sport, 2020) is the *retained EU law* version of the GDPR. Retained EU law is a new category of UK law created by the European Union (Withdrawal) Act 2018 (EUWA).

The Withdrawal Act repealed the European Communities Act 1972 and makes other provisions with regards to the withdrawal of the United Kingdom from the EU (ibidem), de facto retaining the EU law, and making it possible for most of the EU legislation created when the UK was part of the European Communities to continue to apply in the UK.

The UK GDPR and an amended version of the DPA 2018 are now the main DP legislative texts in the UK.

Therefore, the GDPR has been a fundamental legislation in the history of DP in Europe and in the UK. Although Brexit created an initial uncertainty around the willingness of the UK to keep the Regulation, no changes were de facto introduced.

However, this could change soon as the UK government has announced it is looking at modifying the DP regime to promote innovation (UK Department for Digital, Culture, Media & Sport, 2021). Amongst the envisaged measures is the use of AI for providing personalised services. The way the Regulation is referred to in the document and the dichotomic view of innovation and DP show a rather low understanding of the GDPR.

The potential impact of the proposed changes on FAT and the consequences for data subjects are difficult to predict.

2.3.9 Gaps in the literature addressing DP

The review has presented the evolution of the DP right, the new GDPR requirements, and some grey areas regarding the systems that are currently developed. While important issues are raised in the literature, some gaps call for further inquiry. The following questions are barely addressed: How is the GPDR understood and implemented? What are the challenges faced within organisations? How are the requirements concerning Art.22 satisfied? What do organisations expect from the processing done via AI? Are the increased risks linked to the use of personal data, such as profiling or automation, sufficiently considered and assessed? How, by whom, and when are these considerations addressed? Again, as in the case of research into AI, the praxis around the implementation of the Regulation tends to be overlooked, and following that, also the role of people in guaranteeing the responsible use of personal data. This research aims to fill these gaps by looking at the praxis of the GDPR *within its organisational context*. This is done by including the people working with DP and AI, and by exploring what their aims are and what kind of environments and processes they are creating with their decisions.

2.4 FAT principles

This part presents the three FAT principles. After introducing the relation between FAT and AI ethics, the principles are described in detail.

2.4.1 FAT and AI ethics

Interest in the social implications of AI has increased considerably in the last few years. Concerns about the fairness, transparency, and accountability (FAT) of algorithms and ethics around AI are now being discussed more frequently in various fields, raising awareness, and increasing the demand for further research.

Concerns about the consequences of Big Data were first voiced by a report on Big Data commissioned by the Obama Presidency (Podesta et al., 2014), prompting researchers to get together to further explore those issues, firstly with workshops, then with a proper conference (FAT/ML Conference, 2018).

While the focus of the research has been mainly on algorithm design within the domain of computer science, there is a change of direction with the inclusion of other disciplines (Benjamin, 2019; Bird et al., 2016; Eubanks, 2018; Skirpan & Gorelick, 2017). This trend can also be observed in the evolution of the FATML (FAT Conference, 2018; Friedler & Wilson, 2018; Sim et al., 2021), now FAccT (ACM FAccT Conference, 2021), which explores socio-technical systems by including interdisciplinary work (e.g., law and social science).

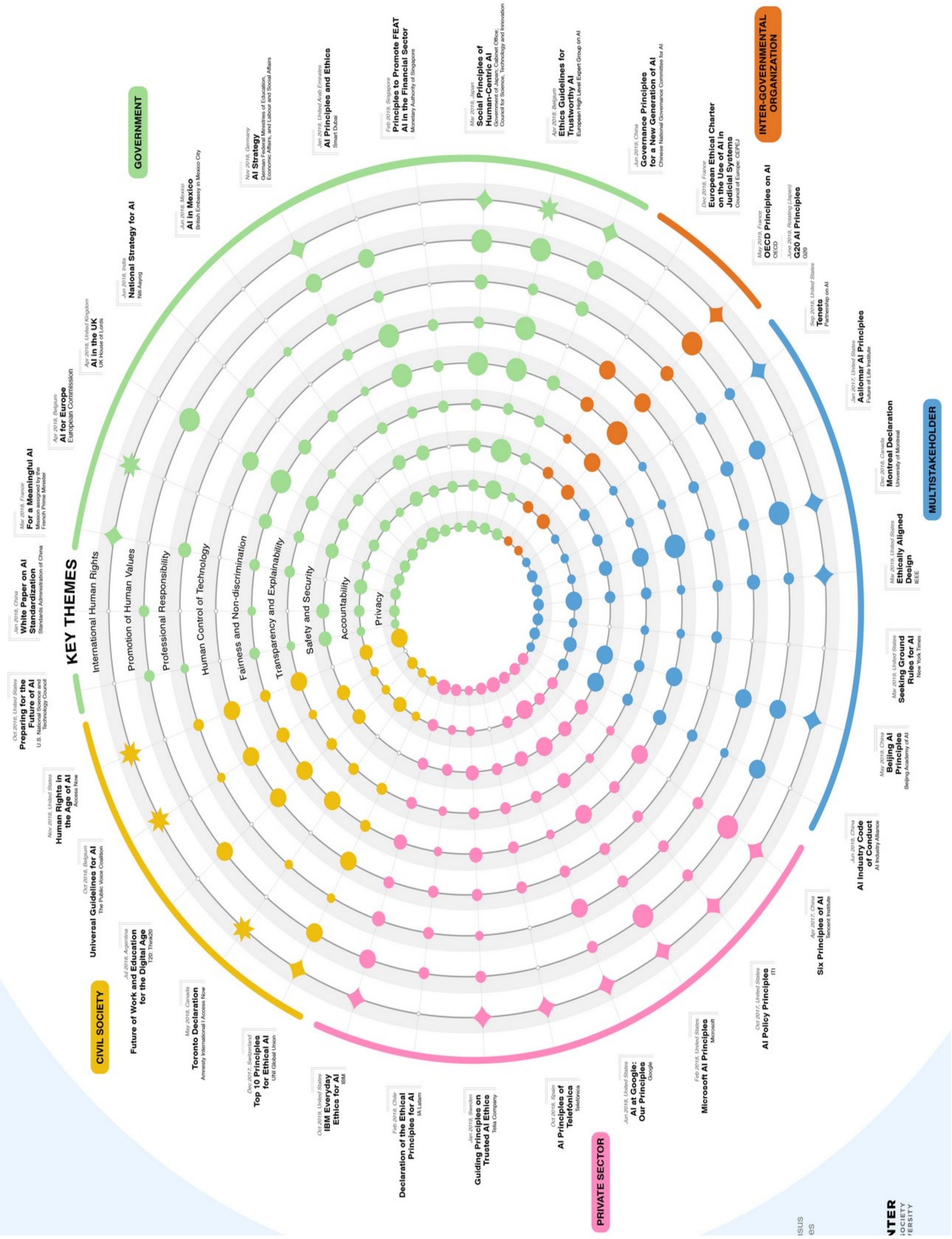
As awareness and debates on AI ethics grow, some organisations and institutions have published lists of principles and guidelines on how to use AI systems ethically, and the number of publications is increasing. For example, Google published a short list of principles (Google, 2018), whose suggestions for automated decisions resemble the GDPR's ones, denoting its impact on corporations. The UK Government issued the Data Ethics Framework (UK Department for Digital, Culture, Media & Sport, 2018). More recently, IEEE Ethical Design (IEEE Board, 2019), Microsoft Ethical Principles (Microsoft, 2019), Organisation for Economic Co-operation and Development (OECD) (2019), GCHQ Report on Ethics and AI (GCHQ, 2021), UK Office for AI guidelines on automated decision-making (Office for AI, 2021). The Principles Artificial Intelligence Project (Fjeld et al., 2020) created by the Berkman Klein Center at Harvard maps the proliferation of AI principles and guidelines (Figure 2-7).

The project included guidelines published by governments, international institutions, companies, and multiple stakeholders, a remarkable document whose interactive version shows the different documents in various domains (civil society, government, inter-government organisations, multi-stakeholders, private sector).

Relevant also is the EU contribution. The guidelines on AI ethics, created by a group of 52 experts and published by the European Commission (High-Level Expert Group on Artificial Intelligence, 2019) recognised both positive impact and risks of AI and saw a trustworthy AI as a combination of three key elements

> (1)…lawful, ensuring compliance with all applicable laws and regulations; (2) …ethical, ensuring adherence to ethical principles and values and (3) …robust, both from a technical and social perspective since to ensure that, even with good intentions, AI systems do not cause any unintentional harm (ibidem, p. 35)

Figure 2-7 AI principles and guidelines



(Source: Fjeld et al., 2020)

The proposal recalls some GDPR requirements, e.g., fines and extraterritoriality. While important in containing some forms of surveillance and predictive technology (i.e., students ' data), the proposal has also been criticised for not regulating some areas, such as automated weapons (B. C. Stahl, 2021a) or not regulating others enough, such as biometrics (Kind, 2021). The future Regulation will not be enforceable in the UK due to Brexit, but it will apply to UK providers selling or providing services in the EU and it is expected to influence future UK legislation on AI.

Similarly, the amount of research around AI ethics has also increased remarkably in the last few years (Coeckelbergh, 2020; Dignum, 2018, 2019; Floridi, 2021; Floridi et al., 2018; Milano et al., 2020; Mittelstadt, 2019; Siau & Wang, 2020) revealing growing concerns but also hopes in the use of AI (B. C. Stahl, 2021b).

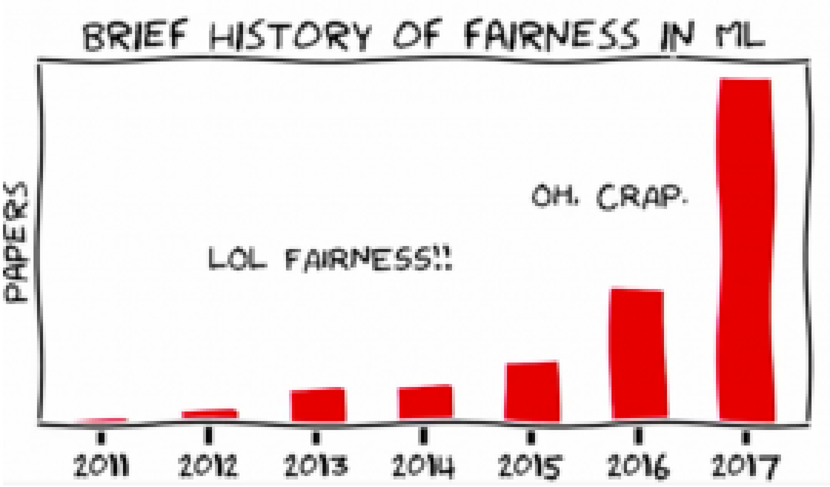
The discourse around AI ethics is wider and complex. This research does not aim at exploring AI and ethics, nor the concept of ethics, the main theories and philosophies. While that would be very inspiring, it is not considered the right approach for this research. Considering how some ethical guidance at times appears to disregard the practical dimension, this research shifts the focus onto that dimension by looking at the FAT principles. Considered as subsets of AI ethics, they are viewed as both closer to the everyday experience of people and as organisational and management issues (thanks to being GDPR requirements). This permits the uncovering of some aspects still largely under-explored in research. For example, the exploration of FAT principles within the implementation of AI, and their adaptability to specific contexts and practical settings.

The main elements in the discourse around FAT and the GDPR are presented below.

2.4.2 Fairness

The debate on fairness, its meaning, and its relation to AI is complex. The concept of fairness is in general linked to non-discriminatory practices or just treatment, and many definitions have been discussed by sociologists, legal scholars, and computer scientists. The current debate on fairness and AI is growing (Figure 2-8) and it is particularly lively, adding complexity to the discussion (Barocas & Selbst, 2016; Bird et al., 2016; Calo, 2017; Crawford, 2016, 2017; Crawford & Schultz, 2014; Hajian & Domingo-Ferrer, 2013; Kamiran et al., 2013; A. D. Selbst et al., 2019).

Figure 2-8 Increase in research on fairness



(Source: Kate Crawford's NIPS 2017 Keynote presentation: Trouble with Bias)

a. Which fairness in which context?

Different definitions of fairness exist. For example, the guidelines on AI ethics (High-Level Expert Group on Artificial Intelligence, 2019) define the concept by identifying two dimensions: substantive and procedural. The first implies a commitment to "ensuring equal and just distribution of both benefits and costs, ensuring that individuals and groups are free from unfair bias, discrimination, and stigmatisation" (ibidem, p. 12), equal opportunity in accessing resources, and AI which does not deceive. The second involves the capacity "to contest and seek effective redress against decisions made by AI systems and by the humans operating them" (p. 13), and links accountability with explicable decision-making processes.

Different meanings of fairness are identified by Narayanan (2018), whose work is relevant in understanding the debate. He identifies 21 definitions of fairness, the arguments behind them, and the limitations of the technical community, considered incapable of understanding the social sense of fairness, continuously evolving. He calls for more involvement of ethics and philosophy scholars, as reality is complex, and it cannot be easily expressed and simplified by algorithm systems. He also identifies some limitations in the concept of fairness in relation to groups, oversimplifications in predictions made via binary classifications, different understandings of the concept by different stakeholders, and lack of flexibility in moving from one concept of fairness to another in automated decision making. Green & Chen (2019) raise another critical point on different perceptions of fairness and responsibility of different actors in automated processing systems. They observe the common opinion on the part of engineers that they do not feel responsible for the consequences of their creations, they see another possible risk in the deployment of automated systems, which requires the urgent need for a moral safeguard: "…the potential for automation bias raises the unsettling specter of situations in which both the engineers developing algorithms and the people using them believe the other to be primarily responsible for the social outcomes" (p. 9). Thus, developers and users can think the other knows better and is responsible for social consequences. Similarly, Packin (2019) considers the trust people can display in AI making fair decisions: "More and more people and institutions are passively outsourcing and then relying on algorithms to make decisions, in order to get more accurate and cost-effective results" (p. 44). The most striking results emerge when algorithms are preferred to human experts, and the impact of such practice on critical thinking:

> People are losing the desire to seek a second opinion, think creatively, compare among options, and actively benefit from their freedom to choose in our democracy. Instead, they rely on algorithms, which despite the halo effect and institutional aura attached to them, are not neutral or objectively accurate (Ibidem)

Such considerations are particularly troubling considering the increasing use of ML to predict future behaviours (without considering the fairness of the result) (Dressel & Farid, 2018), or anticipated to be used instead of senior leaders (Watson et al., 2021). If different people expect others to deal with fairness, or for algorithms to be inherently fair, they do not question what fairness is, or check if systems they are creating or using can produce unfair outcomes, let alone try to solve it. If they believe others "know better" or that algorithms are right, there is not space for critical questions, or for getting second opinions (Packin, 2019).

The need to contextualise the concept of fairness is also highlighted by Van Kleek et al., (2018). A single definition of fairness does not exist, and the statistical "dichotomy" (yes-no) cannot convey the complexity of the concept, which cannot be translated into a machine-readable format and used to make neutral decisions. A similar point is made by Binns (2017), who recognises the difficulty of a unique meaning, adding the risk of generalisation caused by statistical discrimination. "[T]he use of statistical generalisations about groups to infer attributes or future behaviours of members of those groups" (p. 4), and the risk of having algorithms making "generalisation on steroids…(failing) to treat people as individuals" (ibidem), is a clear call for human intervention as per Art. 22 GDPR. Binns critiques a narrow application of fixed categories and indicates the need to consider the contexts in which they are used. Similarly, Skirpan & Gorelick (2017) highlight the need to consider fairness not as a static and general concept but specifically and contextually. Likewise, Webb, Koene, Patel, and Vallejos (2018) invite us to consider fairness in the specific context to reduce the abstractness of the concept.

Furthermore, understanding what is fair can be difficult if patterns in data are mistaken for causation. “The correlations identified by the algorithms point to some type of relation between different data but without necessarily providing an explanation as to what that relation is, nor whether there is a causal link between the data” (Kamarinou, Millard, & Singh, 2016, p. 17). Mistaking a correlation for causation can lead to unfair analysis and decision-making, as illustrated by Caruana et al. (2015) who point out the danger of underestimating the risks while triaging asthma patients with pneumonia at hospitals. Asthma patients ’admission is generally prioritised due to the high risk of pneumonia, and they recover sooner. ML could read this link as causation (asthmatic patient recovers sooner than others, therefore the risk for others is higher), prioritising other patients.

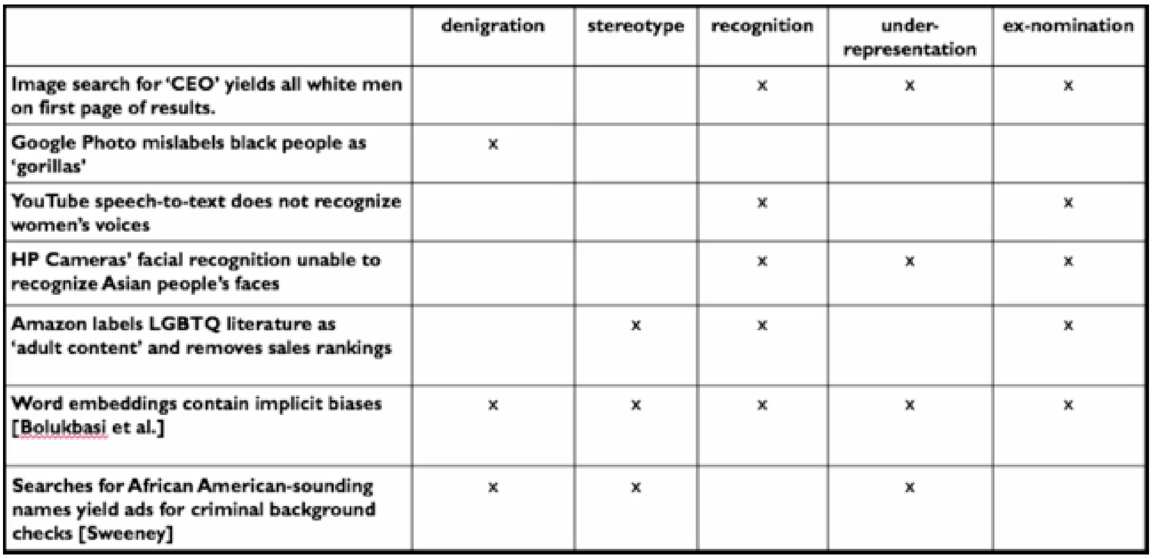
b. Bias

A huge amount of research is now being carried out on AI and bias. Some elements in the growing debate deserve special attention as they can reflect different theoretical positions, (e.g., around origin or responsibility).

Significant is the work by Veale, Binns, & Van Kleek (2018). They argue that unfairness in AI can have different origins, such as sampling of data, fixed social structure applied to different situations and erroneous taxonomies. They question the algorithms’ responsibility in creating unfairness in societies, also asking if the operators should act to reduce unfairness existing in society. Should anti-discrimination be a requirement? "In what way and to what extent? To work this out is a political problem, and one which requires understanding of the context, datasets, and wider social dynamics” (p. 8). Crawford (2016, 2017), one of the most active researchers of fairness and bias, exposed the non-neutrality of data, and the systems of discrimination built into algorithms that can reinforce stereotypes. Analysing the negative consequences of bias, she identifies two different kinds of harms: representational and allocative. Representation harms (Figure 2-9) are the result of systems reinforcing stereotypes, and the oppression of some groups through identity paradigms. For example, Google describing black people as gorillas or systems misreading smiling East Asians as blinking (Crawford, 2016). Representational harms are usually cultural, have social consequences, and diffuse and long-term effects. Allocative harms are the result of a system that fails to provide something to an individual who is part of a group, for example, denying a mortgage to a woman, or considering black offenders as having a higher risk of reoffending (Angwin et al., 2016). These kinds of harms usually have economic consequences, and immediate and short-term effects. Thus, patterns of discrimination can be reproduced by automated processing (Crawford, 2016), creating a "vicious circle of self-fulfilling prophecies" (European Data Protection Supervisor, 2016, p. 4). If getting a loan would depend, for example, on postcode areas, groups that are already oppressed and marginalised could be further discriminated against by the use of biased processes (Rhoen, 2016).

Another important risk identified by Crawford is the emerging trend of ML as a Service (MLaaS) (2017). Biases in MLaaS systems are more difficult to identify and correct, as they are not created in that specific situation/organisation, and they can potentially lack transparency.

Figure 2-9 Example of representational harms



	denigration	stereotype	recognition	under-representation	ex-nomination
Image search for 'CEO' yields all white men on first page of results.			x	x	x
Google Photo mislabels black people as 'gorillas'	x				
YouTube speech-to-text does not recognize women's voices			x		x
HP Cameras' facial recognition unable to recognize Asian people's faces			x	x	x
Amazon labels LGBTQ literature as 'adult content' and removes sales rankings		x	x		x
Word embeddings contain implicit biases [Bolukbasi et al.]	x	x	x	x	x
Searches for African American-sounding names yield ads for criminal background checks [Sweeney]	x	x		x	

(Source: Crawford, 2017)

c. Bias as mirror

Another important distinction in the analysis of bias and fairness, which reflects two different approaches (technical and non-technical), is made by Narayanan (2018) when talking about bias and mirroring. According to Narayanan, some believe that the stereotypes in the real world are amplified by systems, and therefore that representations made by AI systems are not a neutral representation (Hajian & Domingo-Ferrer, 2013; Kay et al., 2015). While others consider the biases in AI to reflect bias outside the system, like mirrors they reflect the real world. According to Narayanan this seems to be the general belief within the tech community, which considers datasets as unbiased samples that can be used for different situations. Representation is considered "neutral", as stereotypes in datasets (e.g., on gender or ethnicity) are found outside the system, and not in the system. The supposed "neutrality" of automated processing was also discussed by Barocas & Selbst (2016), who identified a specific danger. The discrimination could be more difficult to demonstrate if the decision is made via automated mechanism, supposedly neutral, and not via a human process, considered inherently biased. This could create the "perverse result of exacerbating existing inequalities by suggesting that historically disadvantaged groups actually deserve less favorable treatment" (ibidem, p. 674).

d. Which solution?

New solutions are discussed within the technical and other communities. The issues are many, for example, bias identification, neutrality of datasets, more diversity and interdisciplinary approach, debiasing (seen by some as an exercise in applying reductionist thought to complex concepts, Van Kleek et al., 2018), testing, auditing and certifications (Edwards & Veale, 2017), algorithmic impact assessment (Reisman, Schultz, Crawford, & Whittaker et al., 2018), fixing the discrimination acquired through the process of learning (such as the biases learned through actions with humans (Neff & Nagy, 2016).

With the development of AI new questions emerge, and new solutions are suggested by different actors and institutions. These often lack the practical aspects within organisations.

e. Fairness and the GDPR

As previously seen in Section 2.3.6, fairness is included in the first principle of processing personal data: "1. Personal data shall be: (a) processed lawfully, fairly and in a transparent manner in relation to the data subject ('lawfulness, fairness and transparency') (Art. 5.1(a)). Although being a key element in the DP framework, the concept still remains unclear (Clifford & Ausloos, 2017). Guidance published by the ICO (ICO, 2017) on AI, ML and Big Data highlights the risk of creating unfair results for the individuals and suggests organisations to assess transparency and expectations of data subjects while using complex technologies. Butterworth (2018) is critical of the guidance. He argues that "the key challenge for AI processing personal data is in establishing that such processing is fair." (p.

1). As some parts of his analysis are specific to fairness and its limits in predictive profiling, they are particularly significant for this research and will be presented in detail.

1. According to the ICO, unfairness can occur only if a biased dataset is used in the training stage. However, in the case of the COMPAS algorithm (Angwin et al., 2016), used in the U.S. judicial system to support judges, if the decision is challenged, this will be done on a "prediction made about…future behaviour…an opinion rather than a fact… formed by analysing the training data concerning third party individuals, not the data subject." (p. 261).
2. The challenge would be directed at the fairness of training data. However, that would not be the data of the Data Subject and would not be covered by the GDPR.
3. If the decision is not made via automated means the request for human intervention (Art. 22) would not be possible.

Therefore, some predictive profiling "raises questions about the general fairness of the processing" (ibidem) and shows how provisions of the GDPR could be in some cases insufficient in protecting data subjects from unfair decisions. A more recent ICO publication better defines the concept, clarifying that, in DP contexts, fairness "generally means that you should handle personal data in ways that people would reasonably expect and not use it in ways that have unjustified adverse effects on them (ICO, 2020, p. 35).

Veale & Edwards (2018) focus attention on the topics of bias and discrimination, noting how those are mainly non-existent in the Regulation in relation to automated processes. The only reference is in one of the Recitals, which invite the controller to "implement technical and organisational measures…that prevents…inter alia discriminatory effects…on the basis of racial or ethnic origin, political opinion, religion or beliefs, trade union membership, genetic or health status or sexual orientation" (Rec. 71). Therefore, they consider this inclusion of Special Categories of Data to be a sign of discrimination and fairness awareness. "While "fairness" is an overarching principle of the GDPR, it is an extremely under-determined notion in DP that has never been substantially attached to non-discrimination in processing outcomes" (Veale & Edwards, 2018, p. 403). The fact that Recitals are explanatory and not binding, and that Rec. 71 refers only to controllers without mentioning processors and sub-processors, seems to go unnoticed.

Therefore, reality is complex and describing fairness in the real world is difficult. Even more difficult is translating that into a concept to be understood by algorithms. This can lead to a paradox: generalisations can produce simplifications and using AI systems with the aim of making a fair decision could actually generate unfair decisions.

(See 5.5.1 for an example of an AI system created to simplify a complex external reality).

2.4.3 Accountability

As outlined by Koene et al. (2019), accountability can be understood as the complex of practices and mechanisms as part of governance, committed to legal and ethical obligations demonstrating ethical implementation to stakeholders, and remedying any improper act. Accountability and liability of algorithms are being discussed at different levels, and some conflicting views are emerging, signalling the complexity of the topic. Some consider developers responsible and accountable for the harm AI can cause, while others believe that the organisations, or the algorithms, should be held accountable. The EU debated the possibility of giving robots an "electronic" personhood similar to the legal personality of organisations. The more autonomous the systems become through learning, the more difficult it is to consider them as "simple tools in the hands of other actors (such as the manufacturer, the operator, the owner, the user, etc.)" (Delvaux, 2016, p. 6). This proposal faced the strong opposition of more than one hundred AI experts who saw this as a way for producers to escape the responsibility for the harm caused by their products (Robotics Openletter EU, 2018). A more recent publication looks at the effects of a combination of algorithms, data and processes (Koene et al., 2019). The Council of Europe (2017) also indicated some situations where the responsibility is unclear: engineers creating AI systems without knowing the future use and implementation (ibidem, p. 39); organisations implementing AI without knowing how the algorithmic tools operate; ML continuous adapting via learning, such as Reinforcement Learning which is the result of external inputs, continuous interactions, decisions, rewards, and adaptation.

The uncertainty is also extended to the suitability of existing regulations and the need for new ones. The existence of very different opinions in the debate (Ebers, 2019) indicates, once again, the complexity of the issue. While some believe new regulations are urgently needed for AI (Turner, 2018), and the IEEE (The UK Parliament AI Committee, 2017b), others reject this view on the basis that current regulations are sufficient, or that new prescriptions could impact and restrict innovation (Reed, 2018), as for example voiced by DeepMind (The UK Parliament AI Committee, 2017a).

Other issues are data merging (Kamarinou et al., 2016), unpredicted behaviour of algorithms (Tutt, 2017), and decisions made by ML (Kerry et al., 2016). If decisions are made using different sources, the GDPR accountability can be difficult to meet, and the point will be "open to interpretation and need to be resolved in the implementation and interpretation of the GDPR…" (ibidem, 2016).

The principle of accountability in DP is a relatively recent concept, and with the GDPR it becomes the principal basis for legal obligations. The GDPR requests controllers to:

- Be responsible for and demonstrate compliance with the Regulation principles (Art. 5.2).
- Implement "appropriate technical and organisational measures to ensure and to be able to demonstrate that processing is performed in accordance with this Regulation" (Art. 24.1).

As highlighted by the UK Information Commissioner Elizabeth Denham (Denham, 2017), "accountability requires a clear shift in understanding and implementing privacy, and organisations need to understand and mitigate risks " (ibidem), as the GDPR requires the organisation, as a whole, to understand DP and create mechanisms for its application, continuously assess the risks "in all business processes", and demonstrate to the ICO and other authorities how they are putting accountability into practice. This is a call from the ICO for a change in organisations' company culture, which includes DP in risk management, and the understanding of its competitive value, other than its regulatory compliance.

Demonstrating accountability while using opaque algorithms, such as Deep Learning, adds further complexity to the debate, as ML models "are often opaque even to their developers, and releasing the model is unlikely to provide significant transparency" (Koene et al., 2019, p. 30). Butterworth (2018) links the intelligibility of automated processes to the legal personality issue and questions the capability of AI to comply with some GDPR provisions. For example, the "obligations to cooperate with the supervisory authority under Article 31, accountability obligations under Article 5(2) for controllers and information and audit obligations under Article 28(3)(h) for processors." (p. 261).

2.4.4 Transparency

Transparency is strictly linked to explainability. While the first refers to being able to access information about the system, the second indicates the capacity "to understand or explain a system and why it behaves as it does" (Brey et al., 2019, p. 6). Transparency can refer to different factors, such as data (e.g., input, output), algorithms, goals, compliance, influence, usage (Koene et al., 2019).

The existence of a right of explanation in the GDPR has been the topic of "the algorithmic war stories" (Edwards & Veale, 2017, p. 64), and of a lively debate amongst researchers and practitioners on the existence of that right (Casey et al., 2018; Goodman & Flaxman, 2017; Edwards & Veale, 2017; Selbst & Powles, 2017). The new right has been considered and explained very differently by researchers, and this debate is important in understanding transparency in the GDPR and in future research. Goodman and Flaxman see the right of explanation clearly included in the GDPR. Wachter et al. (2017) are more sceptical, and doubt that such a specific right exists in the text, arguing that there is a mandate to provide some information (Art. 13-15) but not a proper right. Selbst and Powles support the existence of the right understood more as "meaningful information" (p. 239). Goodman & Flaxman see in the text a proper "Right of Explanation", the right to have clarification about decisions made through automated processing, a new right that "highlights the pressing importance of human interpretability in algorithm design" (p. 26) and forecasts "a pressing need for effective algorithms which can operate within this new legal framework" (ibidem). Conversely, Wachter et al. reject this argument, arguing that "both the legal existence and the feasibility of such a right" (p. 1) are doubtful, as the GDPR "only mandates that data subjects receive meaningful, but properly limited, information" (Articles 13-15) (ibidem). A third position in the debate is the one by Selbst and Powles (2017) who suggest a more flexible approach, also supported by the different translations of the text in other European languages. They argue for the existence of the right of explanation, but that it is intended as" meaningful information about the logic (involved)…an explanation of some type" (p. 239) that can be found in the purpose and in the text, mainly in Art. 13-15 (Information and Access Personal Data), which "should be interpreted functionally, flexibly, and should, at a minimum, enable a data subject to exercise his or her rights under the GDPR and human rights law" (p. 233).

Deserving special mention in the general debate is the remarkable work carried out by Edwards and Veale (2017). They consider the GDPR provisions unclear and insufficient to trigger “any explanation related-right” (p. 18) and provide the most compelling analysis. Even in the case of a requirement intended as information about the logic behind the processing (ibidem), the kind of explanations that ML systems can provide may be not sufficient to satisfy the request. Therefore, instead of getting lost in the search for a kind of “transparency fallacy” (p. 65), which they envisage could become similar to the search for a “meaningless consent” (p. 23), they question the need to look for a right of explanation, “(I)f meaningful information about the logic of Machine Learning is so hard to provide, how sure are we that explanations are actually an effective remedy and if so, to achieve what?” (ibidem). Therefore, they suggest shifting the focus, from a model-centric explanation, which provides explanations on model, training data, logic, and information about the process of creation of the model, to a Subject-Centric explanation, “which restricts explanations to particular regions of a model around a query” (p. 81). The explanation would not only consider the post-decision stage but would also include a pre-decision and decision-support tool. If only one small part of the system is considered, this can be more interpretable, for example: what changes can I make to my data so that the results could be different? Or what are the characteristics of those who receive a similar outcome? Providing an explanation is less important in Subject-Centric explanations, as the focus moves onto tools, users and subjects that take the decision (p. 58), essentially towards a more Human-Computer Interaction approach. As observed by Edwards and Veale, this constitutes a major shift in the search for a transparent AI, and a change from an “engineering” model (model-centric) to one more based on a Human-Computer Interaction approach. Therefore, they suggest looking at other provisions to create ML systems that are more explicable, for example: Data portability and right to Erasure/to Be Forgotten (not only limited to data provided by the individuals but also extended to data gathered from the observation of their behaviour (online and non), such as likes, or geolocation); Guidance for DPIAs; Certification, privacy seals, and privacy by design, tools able to provide “a more responsible, explicable, and human-centered” (ibidem) answer, more based on a structural process than an individualised one (Crawford & Schultz, 2014).

Casey et al. (2018) add a new perspective to the debate. According to the researchers, most commentators have failed to consider the GDPR’s most revolutionary change, the “new enforcement powers given to European Data Protection authorities (“DPA” in (p. 1), which “provides an unambiguous “Right to Explanation” with comprehensive legal implications for the design, prototyping, field testing, and deployment of automated data processing systems.”

(ibidem). The authors recognise that this holistic approach may not satisfy a strict interpretation of the principle of transparency, intended as a complete and specific explanation of the algorithm. However, they believe the right of explanation to be much wider than that, being related to the adoption and the process of the technology: “legal implications for the design, prototyping, field testing, and deployment of automated data processing systems” (ibidem). Therefore, the interpretation is the result of the audit conducted on the algorithm and data protection by design by Data Protection Authorities, and they predict that this combination of practices will be used as a new framework for entities deploying ML systems. Veale, Binns, & Van Kleek (2018) expand the idea developed by Edwards and Veale of a Human-Computer Interaction (HCI) approach, and they consider how, using HCI to satisfy the GDPR requirements, pre- and post-deployment, organisations can actually improve their governance. As the GDPR is a legislation that has the interaction between systems and people at its core (a proper Information Systems legislation), relying on HCI would not only improve transparency, but it would also improve governance, reduce bias and discrimination, and improve fairness. The connection, made by the authors, between transparency and better governance is of particular importance for this research, especially the reference to “different modes of governance”, as many GDPR provisions “will depend heavily on the communities, practices and technologies that develop around them in different contexts” (Veale, Binns, and Van Kleek, 2018, p. 6): an observation particularly striking which considers the influence of the context on the praxis of DP within organisations.

The last three analyses of the right of explanation appear to be the most compelling and useful in a business environment. A holistic approach to the principle of explanation and transparency, not exclusively focused on the text of the law or on the narrow explicability of the only algorithm, but which considers different tools and privacy as a structural and internal process within organisations, is a promising and valid instrument, considering complexities in organisations resulting from different factors (such as internal systems of power, opaque algorithms, cultural, technical and organisational specificities). Therefore, widening the focus is necessary and real accountability requires a holistic approach involving AI systems, contexts, people and processes.

2.4.5 FAT in organisations

The recent proliferation of work on Ethical AI and FAT is a clear sign of the growing awareness of AI risks, and the desire and need to prepare for it. Yet, some elements are generally missing. For example, processes, verification, and awareness of power and agency in decision making or within organisations are not mentioned. How are those principles put into practice? This was clearly observed by Crawford while commenting on Google AI principles:

> …it's time to ask about governance. How are they implemented? Who decides? There's no mention of process, or people, or how they'll evaluate if a tool is 'beneficial'…Principles minus process, or verification, or internal appeal structure, or independent review = no real accountability (Crawford, 2018)

Crawford also highlights other elements impacting fairness: who decides which issues should be prioritised, or how people should be classified using AI? Such aspects around decision making cannot be omitted, and this is also linked to the GDPR. The new requirements on AI are important in raising the protection of data subjects in the case of decisions made via automated processing. However, they do not regulate all other situations where ML is used to augment human activities, such as performance management, criminal justice (where the decision is taken by judges), or recruitment. Thus, if unfair decisions are taken by humans on top of ML predictions, data subjects will be less protected and potentially more exposed to discriminatory outcomes.

Therefore, in order to understand how the FAT principles can be applied by organisations during the implementation of AI (research question), we need to first understand the main factors in AI management (see 2.5).

2.4.6 Gaps in the literature addressing fairness, accountability and transparency

The review has highlighted how guidance on ethics and debates on FAT are generally focused on the development of AI, disregarding the practical dimension. The question how these principles are translated into concrete practices are under-explored.

This research aims to fill this gap, but in order to do so, we need to understand how FAT are practiced in organisational contexts. Instead of considering the meanings of FAT to be static and given, in this research participants are provided with the possibility to develop their own interpretation of these concepts and to describe how these principles are translated into AI and GDPR-related organisational practices. This approach permits the unveiling of some of the aspects that have so far remained under-explored. For example, the different understandings and the agency of those involved in innovation projects; biases in the process; insufficient clarity in the definition of roles and responsibilities regarding unethical outcomes.

2.5 AI management

2.5.1 AI management: traditional business, innovation, and ML

The last part of the literature review highlights important issues regarding AI management. This focuses on digital innovation, the acquisition and management of AI, some specific challenges of ML to management, and different AI strategies (automation and augmentation).

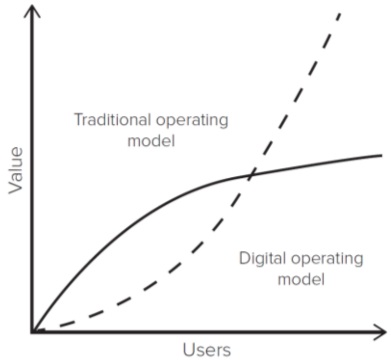
2.5.2 Digital innovation

Digital technologies are disrupting traditional business domains, boundaries between traditional areas are coming down, and new business models are emerging. The virtual economy that has grown in the last few years is, to a stronger extent, based on the distribution of information and on sharing between people than on production in the classical sense (Arthur, 2017). Data, algorithms, and business processes continuously communicate with one another, creating an autonomous "external intelligence which use[s] huge 'libraries' of intelligent functions (that) bit by bit render human activities obsolete" (ibidem, p. 34). According to Arthur, organisations wishing to maintain their competitiveness have two options: create new business models or integrate new technologies into their existing structures. This often requires new innovation strategies that encompass knowledge from different sectors (The Boston Consulting Group, 2018). Significant changes in organisations are needed (Fichman et al., 2014; Verganti et al., 2020) as the more advanced the technologies (such as AI), the more disruptive their effects can be on the business that uses them (The Boston Consulting Group, 2018).

Business dynamics are different in traditional and digital businesses. A low-risk strategy is a common response to business uncertainty in traditional settings. Organisations usually respond by reducing investment, and while a few can experiment, others wait and try later to outperform their competitors (Bughin et al., 2018). Conversely, the pace of digital innovation in digital business is fast, and a new idea can be realised, tested, and changed in a short time. Those who experiment first (and those who follow fast) tend to gain a huge competitive advantage (ibidem). This process is even faster with AI systems. There is a pressing need to embrace innovation: "(t)he need to accelerate innovation and shorten R&D and go-to-market cycles has big implications for how companies manage innovation programs and think about innovation strategy" (The Boston Consulting Group, 2018, p. 9).

As Iansiti & Lakhani (2020) point out, challenges faced by traditional organisations are shared by "the relatively limited impact…on the surrounding economy, environment, and social system" (ibidem, p. 8). At the same time, the growth and impact of digital models and AI firms pose new risks "from privacy to cybersecurity, and from bias to fake news…" (ibidem) (see Figure 2-10), and create new competitive dynamics (Iansiti & Lakhani, 2020b).

Figure 2-10 The collision between traditional and digital operating models



(Source: Iansiti & Lakhani, 2020, p. 8)

The development of innovation in such contexts is particularly interesting for this research: "Development often focuses on producing a minimally viable product, rather than a fully finished version, that companies can launch, collect data on, adapt, and relaunch—all in an iterative, agile style" (The Boston Consulting Group, 2018, p. 10).

Combinations of internal and external data are used by organisations in all innovation processes, and in general, data come from different sources: data from within the organisation, from partners, the data industry, and from interactions on social media (ibidem). Thus, as inactivity could be too risky for business continuity, organisations are using more data (some of which is personal data) and are creating more products and services in a shorter time. This has big implications for DP.

2.5.3 Acquiring and managing AI

The study of AI, its role in digital innovation strategies, its challenges, and its risks for organisations is still in its infancy. Some characteristics in the acquisition of the technology, the management of AI and ML, and its use in decision-making processes, pose specific challenges to DP. They will be analysed below.

a. Acquiring AI and the role of vendors

All Big Tech companies (e.g., Google, Facebook, Amazon, Alibaba, Apple) have invested heavily in AI technologies and, after using them in their operations, are now offering AI services to other organisations for their digital transformation strategies (K. Lee & Ha, 2018). Brey et al. (2019) rightly identify some issues with this trend, as the same companies (e.g., IBM, Amazon Web Service/AWS, Cisco, Microsoft) are now offering "storage, analysis and processing" (ibidem, p. 39) of health data, a practice raising a number of ethical concerns with regard to access to patients 'data.

Smaller organisations are using different ways to acquire, integrate and manage AI (Ammanath et al., 2020; Loucks et al., 2018; Microsoft, 2018), some of which have implications for DP. For example:

1. Enterprise software (Enterprise Resource Planning/ERP and Customer Relationship Management/CRM), which includes AI. Vendors can use an organisation's data to create AI systems, or they can provide ready to use AI systems.
2. Cloud based AI, and especially cloud deep-learning, that can offer "access to immense—and previously costly—computing power necessary to extract insights from unstructured data" (Loucks et al., 2018, p. 5). In this case too, vendors have the possibility of working with organisations to create bespoke AI services.

The researcher sees two potential issues for DP in this context. The use of personal data held by the organisation for creating the system and its periodic updates can give vendors (subjects external to organisations) de facto continuous access to personal data stored by their clients. Furthermore, the use of "out of box" systems provides more accessibility, but the immediate usability by untrained users could provide a false sense of security, potentially leading to a lack of control, thus increasing risk. For example, some algorithms could be created using data taken from other contexts that might not work properly in the new context, leading to non-compliant outcomes in the new environment. As stated by Luca, Kleinberg, & Mullainathan (2016), algorithms are created to make predictions in slightly different situations (i.e., time). Therefore, they are created for "transferring an insight from one context to another" (ibidem). Untrained users would be completely unaware of bias and would not question the system’s quality. The issue is also related to “untrained” management, and this problem cannot be resolved by having vendors in situ for a short time or providing long-distance support. While algorithms can make more precise predictions, they can also “create risks of their own, especially if we do not understand them” (Luca, Kleinberg, & Mullainathan, 2016). Furthermore, in creating bespoke products, vendors are part of the decision on how to process data with AI. The potential influence and power that they have in co-creating AI systems that could be non-compliant (or unfair or inexplicable) seems still to be overlooked in the literature. AI vendors do not merely process data on behalf of the controller but, in many cases, they have an active role in supporting, influencing, co-deciding, and ultimately innovating with their clients. Of relevance is a DPIA on the use of Office 365 published in Nov 2018 (Roosendaal, 2018). The report was conducted on behalf of the Dutch Ministry of Justice and Security and found several issues in Microsoft’s position as a mere processor. However, the possibility of vendors considering themselves as controllers or joint controllers (and therefore liable according to Art. 26 GDPR), and not merely processors, is emerging. Many computer vendors have started questioning their roles as processors, a role considered too prescriptive in the way it has been defined by the GDPR. They are now shifting toward a joint-controller relationship, which would not impact their innovative power. This trend is especially due to the use of AI, as organisations are “(u)sing sophisticated analytics, ML and AI, vendors are increasingly slicing and dicing their customers’ data to do new and clever things, in order to continually evolve their products and identify new services to sell” (P. Lee, 2018).

Therefore, AI brings the promise of massive improvements, mainly via cost reduction and increased performance, and organisations are showing a growing interest in adopting AI systems, specifically ML. As the technology is evolving rapidly, it often involves different stakeholders who are creating and/or using AI. This makes the debate around responsible management more complex.

b. Managing AI

Organisations often do not understand the implications of AI due to a lack of technical knowledge among non-specialist managers (Fountaine et al., 2019; Waters, 2017). Some of the questions debated by leaders and researchers are symptomatic of the uncertainty that AI brings to future organisational structure and management. For example, pressing questions relate to risks, ethics, internal organisation, practical advice for teams who are starting to work with AI, and uncertainties regarding the need to have a business case or an executive sponsor for AI projects (Dyche, 2018). Some of the issues are particularly relevant for a responsible AI management. The following are of paramount importance:

-Business case: Early adopters are reported to be struggling to "articulate a business case or to define success for AI projects" (Loucks et al., 2018, p. 10). This is possibly due to AI being perceived as "experimental" or to the fact that "the group charged with developing an AI solution is unaccustomed to developing business cases to justify its work" (ibidem), which may denote one of the following: a lack of specific strategic planning for AI innovation, the requirement for a more holistic approach involving different subjects and competencies (Ammanath et al., 2020), demanding more integration between teams and areas within the business, and often also some operational changes prior to acquiring the technology (Loucks et al., 2018, p. 10). This mirrors similar organisational needs experienced by organisations prior to implementing the GDPR (Addis & Kutar, 2018).

-Ethics and scientific validity: Research on ethics is mainly focused on bias in data used to create algorithms. Yet, other elements can also create some issues around ethics. Luca, Kleinberg, & Mullainathan (2016) note that often the problem is not the data per se, but "the way we interact with algorithms" (ibidem.): "[M]anagers need to understand what algorithms do well—what questions they answer and what questions they do not" (ibidem). This is crucial when algorithms used in different contexts can lead to unfair decisions due to new biases acquired during the use. Additionally, leaders should also ask "not what AI technologies are capable of doing, but what they should be allowed to do…" (Microsoft, 2018, p. 22). As more organisations are using systems whose scientific validity can be called into question, questions arise around the real capability of AI systems. A growing number of organisations are using ML applications for hiring and performance management. This usually implies biometric tracking, monitoring behaviour of individuals, and questionable interpretation of that data (as seen in 1.2.1). While the implications about surveillance and DP of such applications are being discussed, the aspects around their scientific and ethical legitimacy are often missing. An example is the explanation provided by Humanyze (2018): "[if] every aspect of business is becoming more data-driven, there is no reason why the people side of the business shouldn't do the same" (Areheart & Roberts, 2019, p. 713). Similar views regarding the use of people's data are growing, and the lack of critical discussion on the validity of such views seems to be quite widespread in different industries.

-Human factor: The human factor is more important in AI than in other technologies. As observed by Luca, Kleinberg, & Mullainathan (2016) and emerging in the Loucks et al. (2018) report, management and staff should be informed about the technology and be involved from the beginning. This produces different outcomes: people are informed, involved in the process, have a say, and may understand recommendations and trust it. If people are part of the process and acquire new skills, they are also less scared of losing their jobs, and become an asset for the organisation. Again, this mirrors the involvement of staff necessary for GDPR implementation and compliance.

-Lack of flexibility and clarity: Algorithms are "literal" or single-minded, they are not flexible, and they do exactly what they are told to do. If managers "care about a soft goal, (they) need to state it, define it, and quantify how much it matters" (ibidem, p. 3). Both characteristics can be very important when, for example, a biased decision can be avoided by making specific adjustments.

-Diversity and staff involvement: As previously seen in Section 2.4.2, diversity is a necessary element in reducing bias in algorithms. Diversity in organisations can also help identify bias in the process. This is encouraged by an environment where diversity in opinion and backgrounds is valued, and where different subjects, expertise, areas, and departments are involved in the innovation process.

> (A)doption takes the form of ongoing, iterative improvement, powered by an open, agile culture in which staff analyse and critique the technology as they use it. This allows them to help shape its development based on real experiences, ultimately delivering better outcomes for everyone involved… (Microsoft, 2018, p. 14)

Mixed skill teams are becoming more common. For example, Centrica "made a conscious effort to bring non-technical people from around the business into…the data science team…to provide a real sense of the challenges within the organisation that…(it) should be focusing on solving with AI" (ibidem, p. 20). Diversity should also be sought with regard to the kind of data, the amount and sources. Sources should be "relatively unrelated to one another...This is where extra predictive power comes from...If the data sets are too similar, there won't be much marginal gain from each additional one. But if each data set has a unique perspective, a lot more value is created" (Luca, Kleinberg, & Mullainathan).

c. Specific challenges in managing ML

Managing ML comes with specific challenges regarding data, models, and their use within organisations.

-Data and risks: ML needs three types of data: training data (used to train algorithms); input data (used to make predictions); feedback data (used to learn from the environment where ML is deployed). How output data impacts on the environment is subsequently learned by ML via feedback data. That ML is constantly learning from the context is one of its key features, and yet, it is often forgotten (see CS1). Uszkoreit makes a compelling observation: "[w]e are starting slowly but surely to employ ML in ways where the machine's actions actually have an impact on the world…from which the machine then keeps learning"(Reese, 2018). Noticeable is the lack of consideration for the impact caused by AI-Human interactions.

Data sources for ML can be internal and external. The integration of data from multiple systems can be challenging, i.e., customer data and financial data (Ammanath et al., 2020; Loucks et al., 2018). External data can also be risky, for example, in the case of company acquisitions or mergers not followed by a system integration.

ML systems can also be manipulated using "wrong" data via adversarial data (I. Goodfellow et al., 2018; Loucks et al., 2018) "The adversary can…influence the model to induce incorrect connections between input features and classes (called "false learning") or reduce confidence in the labelling, decreasing model accuracy…the corruption of the training process compromises the integrity of the model" (ibidem, p. 60). Considering the current expansion of biometric data used for authentication, this can have some extremely powerful consequences for data subject rights and political systems.

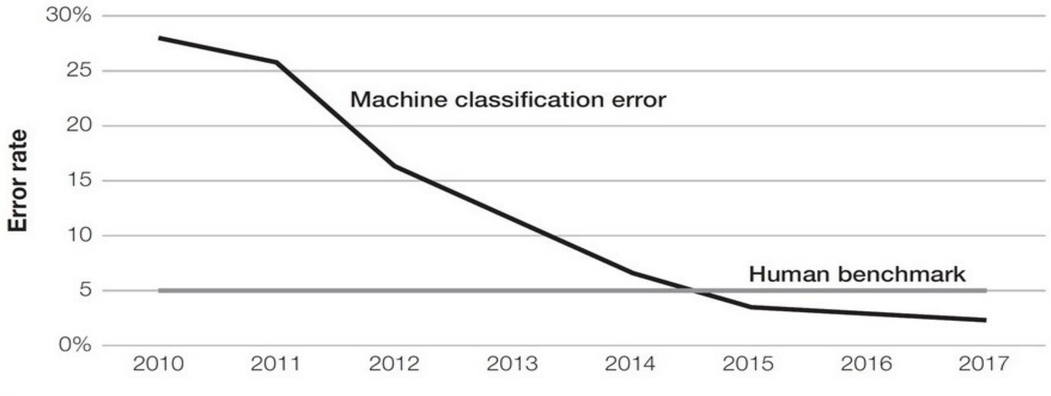
Another issue can be the risk of reverse-engineering the model "by automatically generating large numbers of interactions with a machine-learning-based system, and analyzing the patterns of responses" (Loucks et al., 2018, p. 11). This could make organisations potentially liable for intellectual property theft. Reverse-engineering constitutes a risk for anonymised data as well (Caliskan-Islam et al., 2015).

-Use of ML: In order to understand how ML can be used effectively by organisations, the work conducted by Agrawal et al. (2018, 2019) is significant in providing a clear analysis of ML, what it can do, and its importance for business strategy. They start by observing that what has been so notable in the success of AI is not intelligence, but prediction generated by ML algorithms (as seen in 1.2.2), and by doing so they highlight very specific trade-offs that organisations face while using ML.

> More data means less privacy. More speed means less accuracy. More autonomy means less control. We don't prescribe the best strategy for your business. That's your job. The best strategy for your company or career or country will depend on how you weigh each side of every trade-off (ibidem, p. 4)

Additionally, the more affordable AI, the more pervasive it becomes as new opportunities for use are increasing (Agrawal et al., 2019). While ML can help managers, these have to make significant decisions around the responsible use of ML. As previously seen in 1.2.2, prediction is the act of guessing missing information from available data. Agrawal, Gans, and Goldfarb make a crucial observation. Noticing that the accuracy of prediction increases with the extent of the availability of data, they recognise that in some cases, predictions are so good that "we can use prediction instead of rule-based logic" (ibidem, p. 37). This is a completely different approach to knowledge, and it goes beyond the algorithmic if-then choice. In complex environments where it would be very difficult to label samples, Deep Learning solves this issue with "back propagation" (ibidem, p. 8). Instead of using an IF-THEN algorithmic approach, the predictive approach permits learning via example, similar to the way human brains use memory and experience to make sense of new experience. Figure 2-11 illustrates the improvement of ML in comparison to the human benchmark. This indicates a shift from a deterministic programming of computers to a more probabilistic one.

Figure 2-11 Image classification error over time



(Source: Agrawal, Gans, & Goldfarb, 2018)

However, some characteristics of ML pose specific challenges, such as the following:

-Misreading predictions: Algorithms are extremely good at ”identifying patterns too subtle to be detected by human observation and using those patterns to generate accurate insights and inform better decision making” (Luca, Kleinberg, & Mullainathan). These capabilities are not perfect and considering them as such could be highly risky for organisations. Their probabilistic nature can be easily mistaken for a more deterministic one, and this can create different issues. Some considerations are also significant for DP. ML systems created with personal data are being used to classify and rank people according to various criteria, e.g., physical features, or “moral” qualities, such as honesty (see 1.2.1). Both the scientific validity and fairness of such systems can be difficult to assess and challenge. Misunderstanding the probabilistic nature of ML can lead to dire consequences for data subjects, as it is in the case of predictive justice (Angwin, Larson, Mattu, & Kirchner, 2016).

Furthermore, ML systems are at times seen as being justified on the grounds of a dogmatic scientific assumption that future occurrences will happen exactly like the past, and that, with the right amount of data, the internal qualities of people and their future behaviours can also be predicted. All the other variables which can impact on human behaviour are often not considered at all. This deterministic view has clear implications for business and business models. For instance, new personalised products based on consumers ’DNA and other biometrics are being offered without much scrutiny or deep consideration in relation to scientific validity, DP, and fairness (Hazel & Slobogin, 2018).

-When to use ML. Agrawal, Gans, and Goldfarb also give a detailed analysis of the situations where ML performs better:

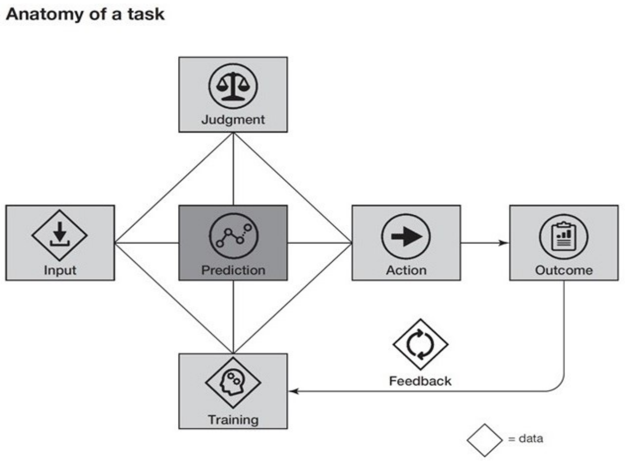
1. Known knowns: High amount of good data makes good predictions, for example, in the case of medical diagnostics or fraud prevention.
2. Known unknowns: Poor or scarce data creates bad predictions, as is the case, for example, with regards to rare past events. Humans perform better than machines in such cases.
3. Unknown unknowns: No data from past events creates poor prediction.
4. Unknown knowns: If the conditions which were correct in the past have now changed, the predictions are wrong, as ML does not understand the change of circumstances.

Humans are better at facing and making sense of new situations, while machines do not perform well with limited data. Humans and machines can provide different kinds of new knowledge, for this reason they can augment each other's performance, perform better and predict better together. Agrawal, Gans, and Goldfarb see this as a new division of cognitive labour, "[t]he human and the machine are good at different aspects of prediction" (ibidem, p. 65). In recalling a case of cancer diagnostics, they note that "[t]he human pathologist was usually right when saying there was cancer…In contrast, the AI was much more accurate when saying the cancer wasn't there" (ibidem). Therefore, while prediction is easy in some situations, and can be performed via automated systems, in others, humans should retain their judgement. Choosing which tasks can be automated is a choice leaders have to make as the consequences for organisations can be significant.

Crucial in the approach of Agrawal, Gans, and Goldfarb is their understanding of the decision-process. They make a clear distinction between predictions and decisions: "making a decision requires applying judgment to a prediction and then acting…humans always perform prediction and judgment together. Now, advances in machine prediction mean that we have to examine the anatomy of a decision." (p. 74). They see decisions as including four different phases:

1. *Prediction* is the combination of training and data, as is the case, for example, if the diagnosis is made by a doctor after considering test results and patient's anamnesis.
2. *Judgement* is made on the basis of a prediction. An example is when a doctor decides the best therapy considering all factors (predictions).
3. *Action* is the decision what to do considering prediction and judgement, such as, for example, receiving therapy.
4. *Outcomes* emerge from action, such as, for example, the relative effectiveness of therapy.

Figure 2-12 Decisions' different phases



(Source: Agrawal, Gans, & Goldfarb, 2018)

Therefore, ML is better at predicting than humans, but prediction is only one element of a more complex process. Humans are still better at performing the other elements of the process. While the authors see the value of prediction going down, they forecast an increase in the value of human judgement. "[A]s prediction becomes better, faster, and cheaper, we'll use more of it to make more decisions, so we'll also need more human judgment and thus the value of human judgment will go up" (ibidem, p. 81). Judgement is the moment of choice, when leaders have to choose between different trade-offs, and in that moment, they could pause and take FAT into account. However, pausing is a cost for organisations, and this has a massive influence on strategic decisions and risk management

Is predicting judgement possible? Complexity and lack of data due to privacy and DP make this impossible as yet.

> As long as enough people keep their sexual activity, financial situation, mental health status, and repugnant thoughts, to themselves, the prediction machines will have insufficient data to predict many types of behavior. In the absence of good data, our understanding of other humans will provide a role for our judgment skills that machines cannot learn to predict (Agrawal, Gans, & Goldfarb, 2018 p. 98)

This is an event unlikely to happen considering the current use of social networks.

2.5.4 Which strategy? Automation to replace or intelligence to augment?

Many organisations are opting for “automation to replace” instead of “intelligence to augment” (Loucks et al., 2018), with AI systems chosen primarily for tactical (e.g., cost reduction) rather than strategic reasons. Companies not using AI to involve staff, augment human intelligence or transform organisations, are missing the full potential of the technology (Wilson & Daugherty, 2018; McKendrick, 2019). While automation will outperform some of the work done by people, AI is seen as able to augment the work done by humans by providing “information, make predictions, and offer alternatives” (Loucks et al., 2018, p. 20).

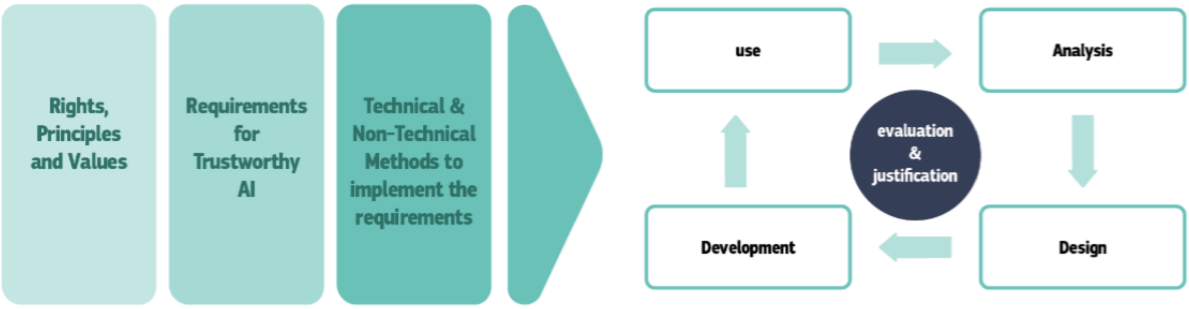
Organisations that include the human factor in their strategies will have to make more changes to succeed. “[H]umans, (will be) using judgment, empathy, and business skill…(t)his is a matter not simply of placing humans in the loop but of the loop being built to augment human decision-making” (ibidem). This is also linked to the culture of the organisation and requires organisations to be focused on key social issues (e.g., data privacy) (Microsoft, 2018).

While both augmentation and automation were chosen by UK organisations (CognitionX, 2018a), the crisis caused by the Covid-19 pandemic has accelerated the adoption of automated systems (Lund et al., 2021). Prior to the crisis, AI was already being used to process staff personal data and special categories of data, especially biometrics. For example, a bank was using Humanyze AI technology to track staff, and measure communication patterns and physical activity to facilitate collaboration between employees (Humanyze, 2018). Humanyze collected data from various sensors, e.g., voices (not the content of speech, but how something is said), movement, and locations. This case already raised a few questions in relation to the anonymisation of biometrics, tracking and consent of data subjects, and real involvement of staff in the implementation and use of the system. Another organisation was collecting and analysing digital footprints of candidates on social media to “determine skills, passions, hobbies, and ultimately, fit with the company’s mission and culture” (CognitionX, 2018b). The Covid crisis has sped up this process, fuelled by the perceived immediate need to check and monitor staff working from home, raising many questions on the processing of staff personal data and special categories of data, especially biometrics (Lund et al., 2021; Trade Union Congress (TUC), 2020).

Being able to clearly understand how to strategically choose the right approach for the right situation can lead to an effective AI management. Additionally, future ethical stances of managers are being discussed, as "AI is right now mainly about making existing processes more efficient rather than having cognitive insights or providing cognitive engagement, but the future appears to be different as AI can have wide effects on the organisation which "may result in significant redesign of workflows and the boundary of the firm" (IESE Business School, 2018, p. 166).

Agrawal, Gans and Goldfarb (2018) advise leaders to be in charge of the strategy and not delegate IT. This approach desires a high level of centralised responsibility that recalls some elements during the implementation of the GDPR (Addis & Kutar, 2018). This would facilitate the realisation of a trustworthy AI which requires technical and non-technical methods in all stages of an AI system (High-Level Expert Group on Artificial Intelligence, 2019).

Figure 2-13 Realising trustworthy AI throughout the system's entire life cycle


Rights, Principles and Values
Requirements for Trustworthy AI
Technical & Non-Technical Methods to implement the requirements
use
Analysis
evaluation & justification
Development
Design

(Source: High-Level Expert Group on Artificial Intelligence, 2019)

Moreover, AI management requires "shifting the emphasis from intelligent systems technologies to intelligent organizations" (Bergstein, 2019, p. 31), "organizations that are enlightened by advanced computing but always cognizant of its limitations and structured accordingly" (ibidem). It is a systemic change that involves all people within the organisation and that will lead to new ways of working. Bergstein is very clear. Intelligent organisations are those that:

> [u]nderstand that at a very fundamental level, accountability and explainability are crucial. People across the organization will have to explain to customers, to each other, and often to regulators, why a decision was made. That will require companies to train employees to see the big picture across the business and to

> understand what data analysis can do and what its limitations are (ibidem, p. 32)

This last sentence is significant. Leaders are facing new challenges, they need to choose very rapidly between different options, and they need to explain those choices.

2.5.5 Gaps in the literature addressing AI management

This review has shown how the debate on the management of AI is mainly focused on algorithms and data. However, to fully understand the complexity of the contexts in which AI is developed, a less reductionist analysis is needed. This research aims to provide such an analysis by including the interactions of humans with the technology, thereby shifting the focus by moving from intelligent systems to intelligent organisations. Within this approach the role of different stakeholders in impacting on the responsible use of AI is considered.

2.6 Conclusion

The literature review explored different aspects of AI, DP, FAT principles, and some aspects of AI management. After describing the evolution of AI, its main applications, and areas of concern, the chapter highlighted some characteristics of ML and some implications for DP. It then moved on to discuss the development of the DP as a right, and how it is legislated within the European and UK contexts. The main focus has been on the GDPR, because the GDPR has significantly changed the landscape of DP in the EU and the UK, with further implications for organisations in other countries. Special attention has been paid to the three FAT principles that have so far not yet been fully embedded in management practices. The final part discussed AI management, defined common challenges and limits in managing AI and ML.

The literature review also identified some gaps in the current research which demand further exploration. The major gaps discussed included:

- A reductionist focus on technology, a lack of research into the application of AI and the organisational context.
- A lack of understanding of the challenges of the implementation of the GDPR regarding the use of AI, the absence of considerations around the specific risks posed by AI, and the lack of attention given to practical and real-life processes.

- A lack of debate around the practical dimension of AI ethics and the FAT principles.
- A scarcity of debate around the human factor in decision making about AI innovation and the interactions of humans with AI within organisations.

By using a holistic approach, this research addresses those issues, and considers the full ecosystem of the organisation in order to understand how different elements and stakeholders contribute to responsible practices. The major argument brought forward by the literature review is that AI, GDPR, and FAT are complex concepts which are not only limited to data management, but which require wider approaches encompassing technology, people, and processes.

CHAPTER 3: METHODOLOGY

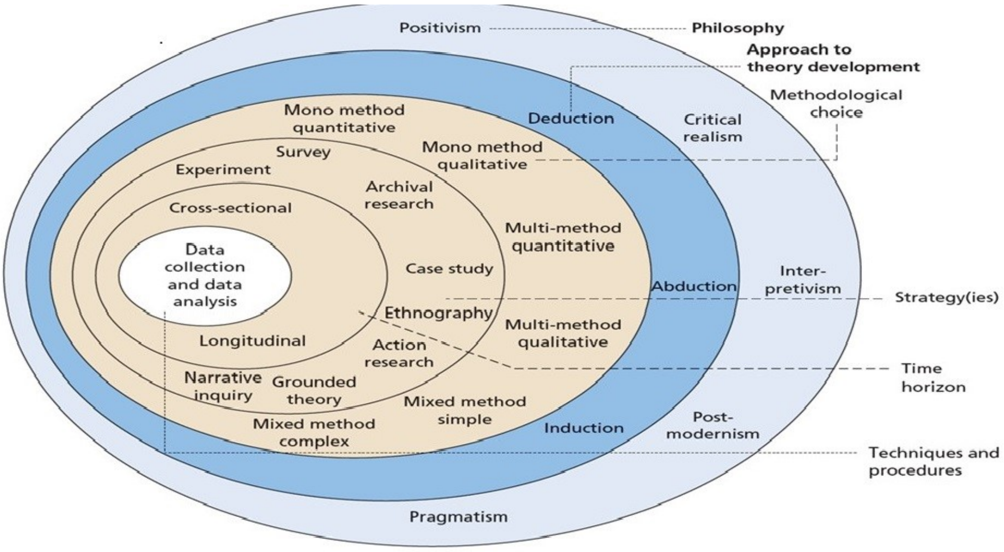
3.1 Introduction

This chapter illustrates the research methodology selected to address the research question. The chapter starts by presenting the chosen research philosophy (interpretivism), research reasoning (inductive), methodological approach (qualitative), and research methods (interviews and case studies). It then presents how participants and case studies were selected, and how data was collected and analysed.

3.2 Research philosophies

This section presents different epistemological approaches and legitimises the stance chosen by the researcher. It starts with a brief reflection on positivism which has for a very long time been the hegemonic approach in social science research (Saunders et al., 2015). The section then introduces the core assumptions within an alternative approach (interpretivism) and explores some of the contradictions within critical realism. The different emphases deployed within these philosophical approaches are well illustrated by the external layers of the research onion (Figure 3-1) created by Saunders et al., 2015.

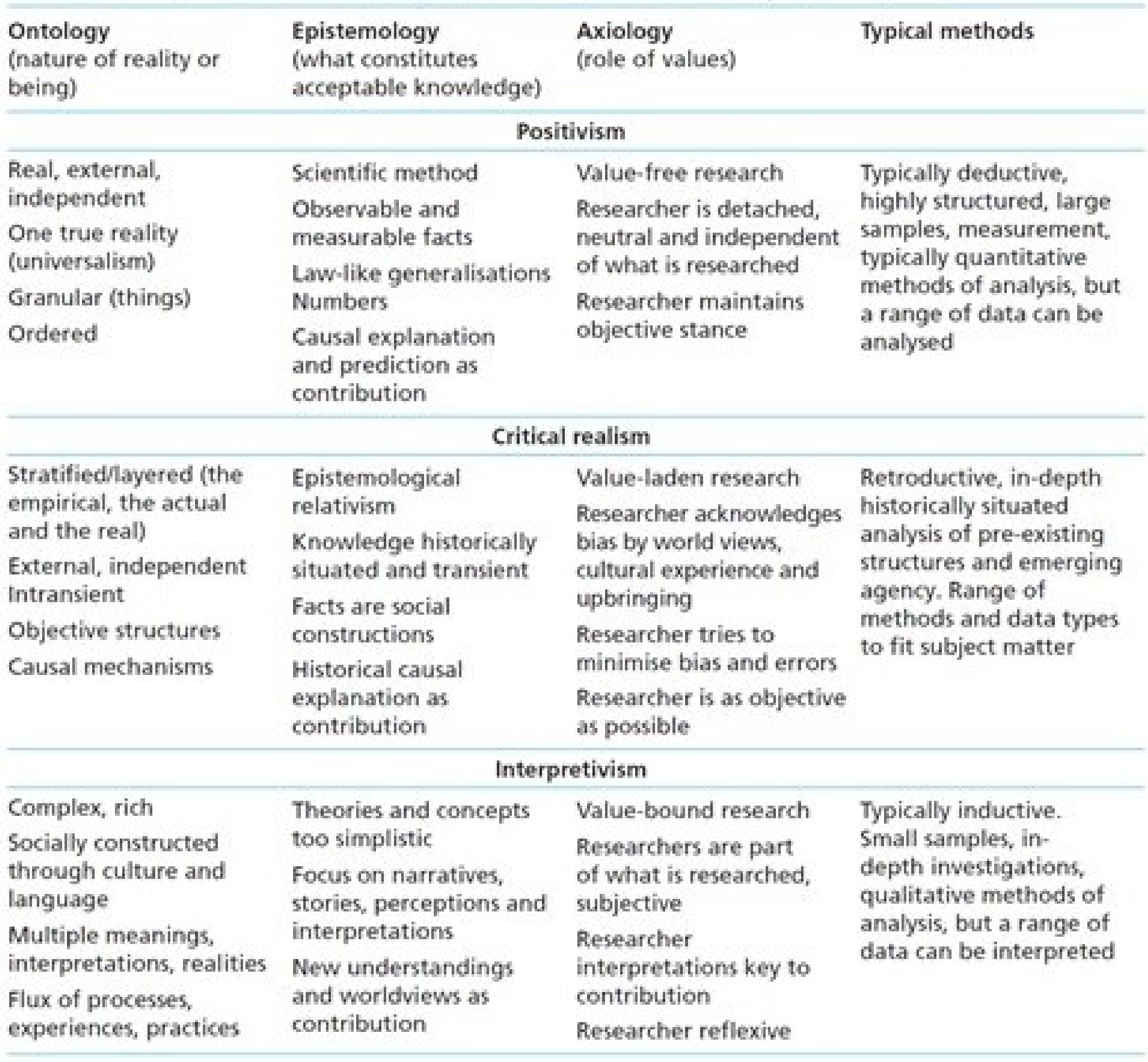
Figure 3-1 The Research Onion



(Source: Saunders et al., 2015, p. 125)

The table below provides a further excellent overview of key differences between the different philosophies.

Table 3-1 Comparison of three research philosophies in business and management research



Ontology (nature of reality or being)	Epistemology (what constitutes acceptable knowledge)	Axiology (role of values)	Typical methods
	Positivism		
Real, external, independent One true reality (universalism) Granular (things) Ordered	Scientific method Observable and measurable facts Law-like generalisations Numbers Causal explanation and prediction as contribution	Value-free research Researcher is detached, neutral and independent of what is researched Researcher maintains objective stance	Typically deductive, highly structured, large samples, measurement, typically quantitative methods of analysis, but a range of data can be analysed
	Critical realism		
Stratified/layered (the empirical, the actual and the real) External, independent Intransient Objective structures Causal mechanisms	Epistemological relativism Knowledge historically situated and transient Facts are social constructions Historical causal explanation as contribution	Value-laden research Researcher acknowledges bias by world views, cultural experience and upbringing Researcher tries to minimise bias and errors Researcher is as objective as possible	Retroductive, in-depth historically situated analysis of pre-existing structures and emerging agency. Range of methods and data types to fit subject matter
	Interpretivism		
Complex, rich Socially constructed through culture and language Multiple meanings, interpretations, realities Flux of processes, experiences, practices	Theories and concepts too simplistic Focus on narratives, stories, perceptions and interpretations New understandings and worldviews as contribution	Value-bound research Researchers are part of what is researched, subjective Researcher interpretations key to contribution Researcher reflexive	Typically inductive. Small samples, in-depth investigations, qualitative methods of analysis, but a range of data can be interpreted

(Source: Saunders et al., 2015)

In the following, the researcher will briefly reflect on pros and cons of these three approaches in order to explain the choice of an interpretivist approach in this research project.

3.2.1 Positivism

Positivism is the research philosophy that assumes an objective reality that is observable, measurable, and can be understood through empirical evidence using logic and reason (Crotty, 1998; Macionis & Gerber, 1997; Saunders et al., 2015).

It originated from the work of Auguste Comte (Mill, 1887), Francis Bacon, and the ”First Vienna Circle”, later called neo or logical positivism (Uebel, 2006). It is based on the assumption that scientific facts and interpretations of data are unambiguous and not dependent on mutable opinions of researchers. Researchers develop hypotheses from theories that already exist and test them by staying neutral and detached (Crotty, 1998). Positivism excludes ethics, aesthetics, culture and religion, which are considered to be cognitively "meaningless—nonsense" (ibidem, p. 26). Positivist philosophy is deeply rooted in scientific knowledge, for example, by viewing scientific results as independent of the positionality of researchers (Hacking, 1981). However, some limitations in using a positivist approach to read AI have emerged, and this is especially the case with Ethical AI. Positivism goes hand in hand with a technological rationality and pragmatism that tends to support easy and one-dimensional answers to ethical questions. For example, AI Now (Whittaker et al., 2018) highlights "(t)he limits of technological solutions to problems of fairness, bias, and discrimination" (ibidem p. 7). Narayanan (2018) identifies the limitations of the technical community in understanding the social meaning of fairness, that cannot be easily expressed and simplified by algorithm systems used to make neutral decisions.

Following this line of reasoning, the researcher does not consider positivism suitable to express the complexity of the research topics and the most recent developments explored in this research. A more complex and humanistic methodological approach is required.

In the following, the benefits and potential shortfalls of a number of interpretivist approaches will be explored.

3.2.2 Interpretivism

Interpretivism opposes the scientific interpretation of reality as measurable given by positivism. Based on the idea that human and physical worlds are different and separated, subjects are assumed to observe the world, create meaning, and use it to interpret the physical world. Researchers are not passive observers, but part of the observed world seen from an informant's point of view (Saunders et al., 2015). Phenomena are both interpreted in their environments and said to be affected by these interpretations (ibidem). The coexistence of multiple cultures and the intersection of experiences in contemporary societies demand the consideration of contexts as an essential part of any philosophical approach that claims to be able to read and understand this complexity. Interpretivism does not share the positivist assumption of an existence of universal laws. Thus, multiple readings of reality become part of the observed reality.

Interpretivism originated mainly from the work carried out within modern hermeneutics, phenomenology, and symbolic interactionism (Crotty, 1998). Hermeneutics analyses events via the interpretation given by social actors, who are conceived as acting and sharing reality with others. This is done considering the social contexts where events occur, contexts which therefore become critical in reading subjective interpretations of phenomena and personal experiences. Prominent hermeneutic philosophers were Heidegger, who considered cultural and social contexts as an essential part of the reality seen as a whole (Heidegger, 1988), and Gadamer, who considered the role of deeply rooted pre-judgements in understanding reality (Gadamer, 2006).

Phenomenology is an approach that studies reality not as something that is separated from individuals, but as lived experiences. The most prominent phenomenological philosopher was Husserl, who was concerned with the structures of consciousness and interactions: "subjects who are not simply reacting automatically to external stimuli, but rather are responding to their perception of what these stimuli mean" (Laverty, 2003, p. 22).

Alternately, symbolic interactionism believes human interaction to be central in societies, and that shared understandings and processes of interpreting those interactions create and recreate meanings (Blumer, 1986; Mead, 1934).

Some recent developments are particularly interesting in relation to AI. For example, hermeneutics is used to understand human interaction with AI systems (Zhu & Harrell, 2009). Digital hermeneutics questions "…the interpretational autonomy of human beings…[the] general loss of control on the way we interpret, and hence see the world" (Romele, 2019, p. 21). Phenomenology of cognition is helping to understand what is needed to duplicate human intelligence in machines (Beavers, 2002), while symbolic interactionism is helping to design Distributed Artificial Intelligence (DAI) in ML (Strübing, 1998).

This research will build upon such work and add to the growing body of research into AI from within an interpretivist framework.

The next section will briefly touch upon critical realism, which was considered as a potentially promising framework for this study and introduces the reasons for not adopting this approach.

3.2.3 Critical realism

Combining scientific realism and social science, critical realism aims at exploring critically the experience of reality, its perceptions, representations, and underlying structures (Saunders et al., 2015). Critical realism originated from some critical positions (such as transcendent realism and critical naturalism) within dominant positivist understandings of science in the western philosophy of science (Archer et al., 2013; Bhaskar, 2013). It is closely linked with the work of Roy Bhaskar, who posits that "facts can entail values" (Collier, 1999), that there is a distinction between the real and the trans-factual, and that there is a difference between "pure and applied sciences and explanations" (Bhaskar, 2013, p. 5).

Critical realism aims at providing explanations by looking at understanding the underlying causes of events, taking into consideration the big picture of historical, social, and organisational structures and their evolution over time (Saunders et al., 2015).

Therefore, there are main dimensions, the subject's personal experience of events, and social constraints and limitations given by political and social situations. Thus, differently from positivism, in critical realism individuals 'understanding is dependent and informed by external structures, that are, at the same time, dependent on people's agency and resources.

Critical realism has been the subject of specific criticism. For example, Alvesson and Deetz (2000) argue the approach is too theoretical and lacks pragmatism. Hammersley (2009) suggests that many researchers within the tradition "…fail to explicate the basis for their critical orientation…(and that) social institutions are presented as if their undesirability, and the need to change them in particular ways, were immediately obvious" (ibidem, p. 2). Being critical per se is seen by Hammersley as an aprioristic assumption that bears judgment, therefore critical realist researchers already assume the validity of their assessments of the condition of society.

The researcher finds Hammersley's position particularly convincing, especially:

1. The preeminent emphasis on criticism in critical realism is potentially able to increase bias. "(T)he simultaneous attempt both to produce knowledge and to bring about a social change of some kind (or, for that matter, to preserve the status quo) is liable considerably to increase the danger of bias" (ibidem p. 7).

2. His denial that social scientists, independently from being realist or not, do not possess any special insight in understanding what is good or bad in the situations they

> are exploring, and that “they should not pretend to have this capacity” (p. 8), even “where value judgements rely on research evidence they also necessarily depend upon other factual assumptions and upon value principles that are plural and often in conflict*”* (ibidem).

Even though critical realism could appear to be, at first glance, an appropriate philosophy for this research (because of its emphasis on structures as a conditionality for human action), the researcher does not believe it can provide a consistent and non-biased alternative to interpretivism.

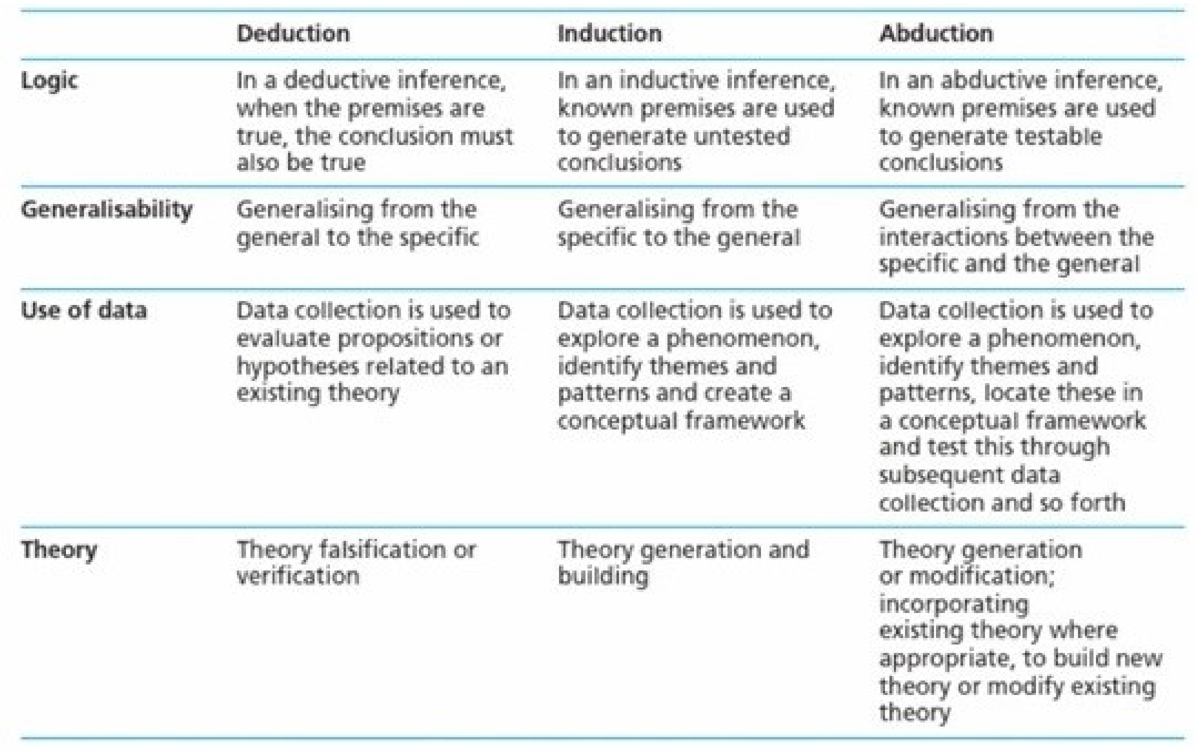
This research investigates how different subjects within organisations understand and use AI and personal data within a DP regime, and how their interpretations may make certain events happen. It is appropriate to assume that different subjects offer different explanations and hold different views on the social and organisational context. The researcher takes the view that people’s opinions can shape the topics of this research, considering the ambivalent nature of AI, and its multiple interpretations and societal concerns. Because of the stronger emphasis the researcher places on people’s agency, interpretivism appears to be a more appropriate philosophical framework for this study. Arguably, interpretivism prioritises the human subject, whereas critical realism tends to prioritise social structures (in the sense that structure always precedes agency).

Additionally, this research draws upon the critical theory tradition (see Chapter 7), which is considered to be compatible with interpretivism, commonly applied in IS research (Sthal, 2008). This approach can be referred to as ‘critical interpretivism’. As argued by Pozzebon (2004), interpretivist approaches explore how a specific social reality is constructed, while critical ones focus on how dynamics of power and ideologies shape those social practices. As the boundaries between interpretative and critical are considered as a matter of degree, they are not necessarily incompatible. By understanding research contexts from the interpretations and the reflection of power, IS research can be both “interpretive and critical without any inherent inconsistency” (ibidem, p 278). Thus, the boundaries between interpretative and critical are fluid, allowing for mutual alignment and/or overlaps.

3.3 Research reasoning

The process of scientific reasoning can be based on three different approaches: deductive, inductive, and abductive. The core assumptions bound up with these different modes of reasoning are well described in Table 3-2. Deductive reasoning goes from general to specific, and from premise to conclusion. Deductive reasoning, also called the Aristotle method, was firstly introduced by Aristotle in his *Prior Analytics* (Smith, 1989). He presented the syllogism as a form of logical reasoning leading to a conclusion that follows from two premises (major and minor), both supposed to be true (e.g., all men are mortal, Socrates is a man, therefore Socrates is mortal). In deductive thinking "…the conclusion must follow analytically from the premises; the normative rule for reasoning is logical coherence" (Ketokivi & Mantere, 2010, p. 5).

Table 3-2 Deduction, induction, and abduction: from reason to research



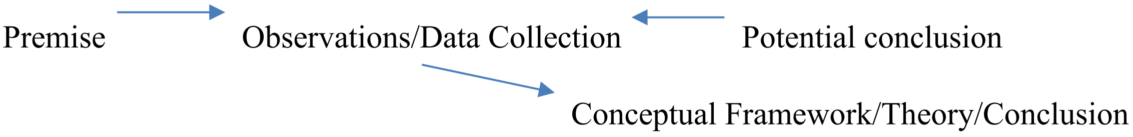
	Deduction	Induction	Abduction
Logic	In a deductive inference, when the premises are true, the conclusion must also be true	In an inductive inference, known premises are used to generate untested conclusions	In an abductive inference, known premises are used to generate testable conclusions
Generalisability	Generalising from the general to the specific	Generalising from the specific to the general	Generalising from the interactions between the specific and the general
Use of data	Data collection is used to evaluate propositions or hypotheses related to an existing theory	Data collection is used to explore a phenomenon, identify themes and patterns and create a conceptual framework	Data collection is used to explore a phenomenon, identify themes and patterns, locate these in a conceptual framework and test this through subsequent data collection and so forth
Theory	Theory falsification or verification	Theory generation and building	Theory generation or modification; incorporating existing theory where appropriate, to build new theory or modify existing theory

(Source: Saunders et al., 2015, p. 146)

Thus, the conclusion is true when all premises are true (ibidem). Confirming the premises, the conclusions do not include or provide new knowledge. Therefore, by adopting deductive thinking, researchers formulate hypotheses within specific theories and collect data to prove them:

Theory/Premises → Hypothesis → Testing Hypothesis → Conclusion

The inductive approach is a form of reasoning that develops and goes from specific cases to general rules. Originated from the criticism of deductive thinking, the inductive approach assumes the existence of a "logical gap" between premises and conclusion and produces different variants of inductive reasoning (Ketokivi & Mantere, 2016). For example, the Baconian method (elaborated by Francis Bacon) considers the beginning of the search to consist in "pure" observation and states concerned with the need to have an intermediate hypothesis in between premises and conclusions (ibidem). Hume (2009) notes the role of the subject's experiences in generating knowledge, and questions the universal truth of scientific phenomena, and the role of opinions, beliefs, and habits in creating scientific values. Saunders et al. (2015) highlight the existence of a knowledge gap between premises and conclusion. Premises give some indications that the conclusions could be true. Collection and analysing of data could lead to the conclusion:



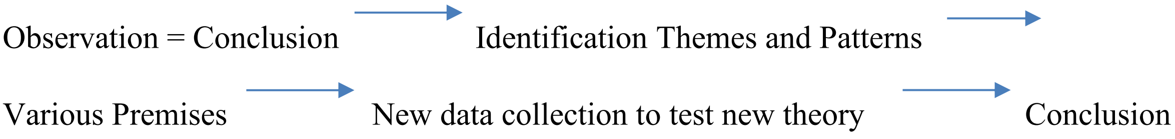
Premise → Observations/Data Collection ← Potential conclusion

→ Conceptual Framework/Theory/Conclusion

Inductive reasoning is usually related to in-depth qualitative analysis of small samples (Saunders et al., 2015) used when deductive reasoning is not appropriate, for example, when it is important to understand the insiders 'perspectives.

Another approach to research design is based on abductive reasoning, i.e., a combination of deductive and inductive reasoning, strongly linked to inductive reasoning (Bell et al., 2018). The most famous version of abductive reasoning is also called "inference to the best explanation" (Lipton, 2003). Subjects are always inferring, explaining, and creating new theories in order to make sense of reality: "we work out what to infer from our evidence by thinking about what would explain that evidence" (ibidem, ix).

Abductive reasoning begins with observing "a 'surprising fact'" (Ketokivi and Mantere as cited by Saunders, Lewis, & Thornhill, p. 144), or a set of facts, which are considered as a conclusion. Various premises (themes and patterns in data and other theories) are then taken into consideration in order to provide a possible justification (often with an iteration between theory and data). The best explanation is then chosen, with a conceptual framework, conclusion, or a new theory, which will eventually modify existing theories:



Observation = Conclusion → Identification Themes and Patterns →

Various Premises → New data collection to test new theory → Conclusion

This research project does not aim to test hypotheses that are part of existing theories (as per the deductive approach), but aims to understand why particular phenomena are happening, focusing on the role of subjects who are working within organisations, their thoughts, experiences, and how these are informing the use of AI, all of which can lead to generating untested conclusions. Furthermore, deduction can be highly structured, or can lack flexibility, and not allow alternative hypotheses. This means that abductive reasoning, too, due to its reliance on a deductive element is not appropriate either. As highlighted by Saunders et al, abductive reasoning is more appropriate to describe a phenomenon, while induction is more suitable to understand the reason for the occurrence. Induction is flexible, considers specific contexts and perceptions of interviewees, and is regarded as more suitable for researching small samples (e.g., case studies) and exploring innovative topics.

Therefore, the researcher considers the inductive approach to be the most appropriate for this research. By using inductive reasoning, the researcher aims to analyse data collected through interviews and case studies and to produce some general untested conclusions and hypotheses on the use of AI and its implications for DP.

3.4 Methodological approach

The process of understanding social phenomena through the analysis of data can be performed using qualitative, quantitative, or a mixed method approach. Each method aims at answering different research questions.

Quantitative methods analyse and measure relationships amongst variables, aiming at describing reality. They are utilised in the case of research that starts with a problem, is based on objective measurements of quantified data, and is used for verifying hypotheses, conducting experiments, and replicating phenomena. Standard techniques, unique and objective interpretation of phenomena, and replicability of experiments are some of the characteristics of quantitative methods (Adams et al., 2007; Bell et al., 2018).

Saunders et al. (2015) argue that there is a connection between research philosophy, research approach, and quantitative methods. Quantitative methods are generally linked to positivist philosophy and are used to verify hypotheses, even though they can serve to elaborate theories. Conversely, qualitative methods are linked to a “holistic approach that involves discovery” (Williams, 2007, p. 67). Based on inductive thinking, they aim at understanding complex social phenomena which need to be explored and unfolded (ibidem). The strong correlation between observer and data is a key difference if compared to quantitative research. Researchers are not passive observers, and in describing and interpreting data they are active elements and thus part of the investigation.

Mixed methods are a combination of quantitative and qualitative strategies. Complementing each other, mixed methods are characterised by "…methodological pluralism or eclecticism, which frequently results in superior research (compared to mono method research)" (Johnson, Onwuegbuzie, & Turner, 2007, p. 14). Key elements are the collection and analysis of both qualitative and quantitative data, use of rigorous qualitative and quantitative procedure, and integration and combination of qualitative and quantitative data (Creswell, 2014).

As this research is aimed at understanding the impact of new events, such as the impact of AI technologies and new legislation on people and organisations dealing with innovation, it is advisable to undertake explorative research using qualitative methods, which can provide answers to complex contemporary phenomena (by addressing what, how, and why questions).

Furthermore, awareness and exact meaning of some concepts (e.g., fairness) are deeply shaped by the beliefs, interactions, and experiences of individuals who are already using the technology. Additionally, implementations and regulations can be very different across sectors and industries, and new trends, such as new geopolitics, or new forms of surveillance fuelled by the Covid-19 pandemic, contribute to creating a complex picture. What is needed is an approach that explores, in detail, how subjects who have agency within organisations are reacting to internal and external events.

Therefore, this research uses qualitative methods, as quantitative methods were not considered sufficient for capturing and understanding the complexity of the topics.

3.5 Research methods

Interviews and case studies were the chosen research collection methods for this project. The main characteristics of these methods and the reasons for their selection are explained below.

3.5.1 Interviews

Semi-structured interviews were chosen to enable participants to answer by giving more detailed responses, thus having the potential to identify new ways of seeing and understanding the research questions (Bowling, 2002; Saunders et al., 2015). Semi-structured interviews allow collecting detailed data with more flexibility, as the question can be adapted according to different sectors. Furthermore, researchers can ascribe significance to a respondent's silence and adapt or change questions according to specific situations.

3.5.2 Case studies

The main aspects of the case studies will be presented below.

a. Definitions

There is not a standard definition of a case study. Robson (2011) describes it as "…an empirical investigation of a particular contemporary phenomenon within its real-life context using multiple sources of evidence" (p. 146), while Yin (2003, 2014, 2018) provides a more complex definition, identifying two parts related to the scope and mode of enquiry of case studies:

1. Scope. "(A) case study is an empirical method that investigates a contemporary phenomenon (the case) in depth and within its real-world context, especially when the boundaries between phenomenon and context may not be clearly evident" (2018, p. 15). Yin highlights phenomena, contemporaneity, and context (similarly to Robson), but he also adds boundaries and stronger focus on context as essential parts of a case study.

2. Inquiry. Yin highlights the multiplicity and importance of theoretical propositions. A case study inquiry copes with a situation where "there are more variables of interest than data points…and benefits from the prior development of theoretical propositions, to guide design, data collection and analysis, and…relies on multiple sources of evidence, with data needing to converge in a triangulating fashion" (ibidem).

Thus, case studies aim at understanding contemporary phenomena using data gathered from multiple sources in their contexts (and not in laboratories), and at understanding of internal processes, outcomes, and effectiveness of events.

b. Types

Yin identifies three types of case studies:

- Exploratory, aimed at developing hypotheses and propositions, usually used in case of scarce existing knowledge, which forms the basis for a preliminary understanding.

- Descriptive, which describes phenomena, context, and impact. The analysis describes an issue, or processes, what has been done and how, how it was perceived by different people, and its outcome.
- Explanatory, which explains how something has been used, how situations and outcomes happened, how different elements are inter-linked, or how concrete situations explain the theory.

A further type, demonstration, is identified by Lazar (2017) who discusses case studies carried out within Human-Computer Interaction (HCI).

Case studies can also be differentiated in relation to their respective approaches. For example, their approach to time, with historical, short-term/contemporary, and longitudinal study (investigation done in different moments), or their approach to theory, as these can serve to build a new theory, test an existing one, or "see which pre-existing theory or model best matches what was found in the case" (Robson, p. 146).

This research is likely to have elements of each type. It is a contemporary study and aims at building new models.

c. Characteristics

Case studies have specific characteristics which are clearly listed by Benbasat et al., (1987, p. 371). Contemporary phenomena are observed in their natural settings, and not recreated in laboratories. They do not involve experiment, but exploration, classification, and hypothesis development. The data is gathered via multiple sources and means, and the change of data collection methods is possible. Case studies are useful to explore the motive (why) and process (how) of events, but they are dependent on the capacity of the researcher to integrate various parts (e.g., through triangulation).

The use of multiple data sources and methods, such as interviews, questionnaires, document analysis, archival records, focus groups, and observations, helps to triangulate and to deeply understand the research contexts (Bass et al., 2018). The capacity for producing complex explanations make case studies a widely used method to understand reasons and processes in business and management research.

However, some criticisms also exist, for example, regarding the unreliability of self-reporting data (which makes using multiples sources, especially observations, particularly important), unsubstantiated observations (especially at the beginning of the observations), and unsystematic summaries. The risk of unwarranted generalisation of the findings and their uncritical application to other settings is a common criticism of case studies (Gibbs, 2012). While some elements of the case can be unique and therefore not generalisable, some authors, such as Oates (2005) recognise the common existence of some elements that are not necessarily specific (such as location, sector, or type of organisation). Therefore, while some elements may be unique, others may be generalisable and transposable to similar scenarios.

d. Case study design

Case studies can be designed as a single case, which is the most common approach (Oates, 2005), or multiple cases (Stake, 2013) (see Figure 3-2). Single cases are appropriate when five rationales occur (critical, unusual, common, revelatory, longitudinal), but they have some limits. For example, single cases can be risky if they later turn out to differ from the original design (Yin, 2018, p. 49), and they can also provide a weaker basis for the generalisation of findings, as is the case with systematic comparison in cross-case analysis that identifies similarities and missing differences (Ridder, 2017). Multiple cases are generally considered to be more robust, as the evidence is more compelling (Herriott & Firestone, 1983), similarities and differences are evaluated, and some elements from one case can be tested or help to explain the second case (Oates, 2005, p. 144). A further distinction made by Yin is between holistic and embedded case study design (see Figure 3-2).

When various levels of analysis are needed, the design should include the embedded sub-units within the case. However, if no multiple levels are required and only the analysis of a single aspect is deemed sufficient, a holistic design can be used. Both designs bear some risks, e.g., too much focus on sub-units or a single aspect, an excessive degree of abstraction, or the risk that some changes can occur during data collection. For example, if researchers are not aware of slippages, have not considered them in advance, or are not prepared for a contingency plan, these factors can impact on the results.

Figure 3-2 Single and multiple-case, and holistic and embedded designs

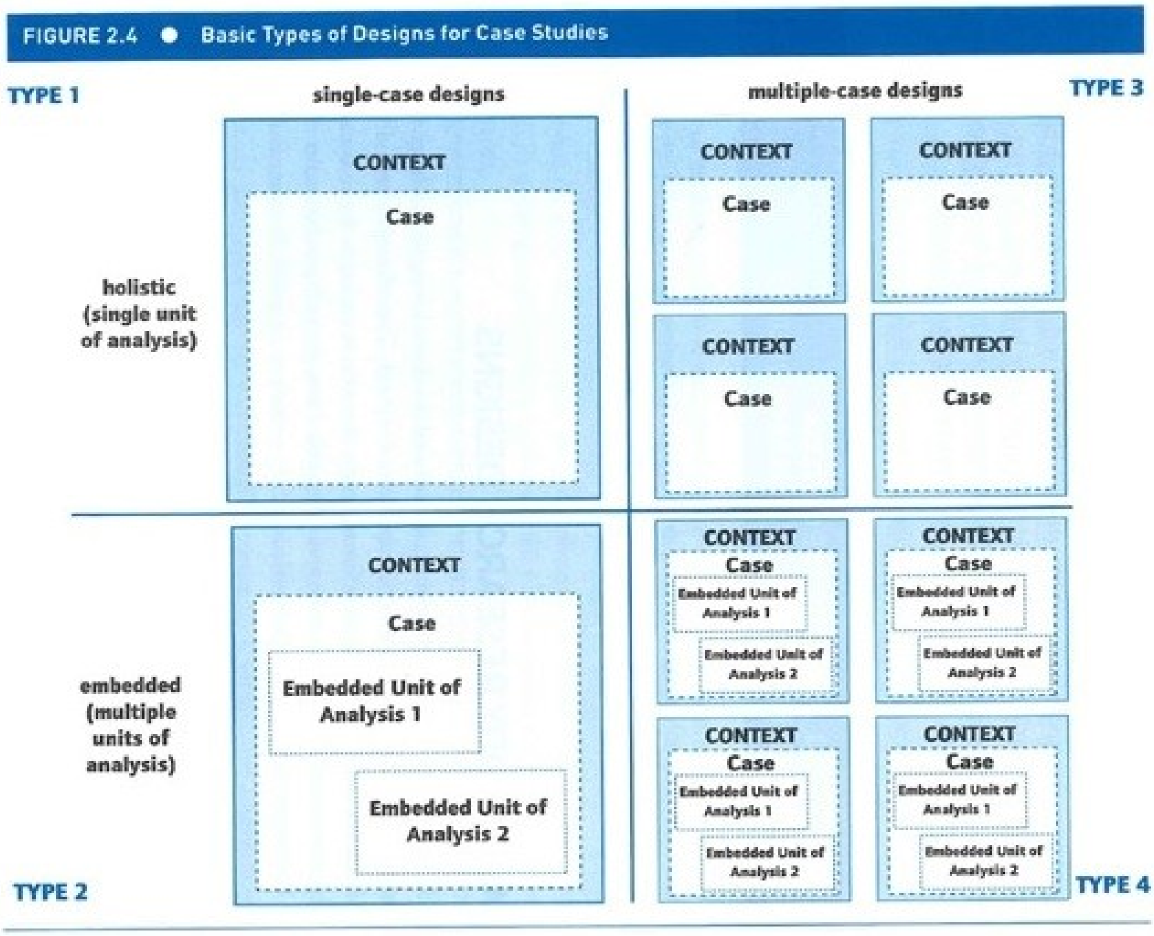


FIGURE 2.4 ● Basic Types of Designs for Case Studies
TYPE 1
single-case designs
multiple-case designs
TYPE 3
CONTEXT
Case
holistic (single unit of analysis)
embedded (multiple units of analysis)
Embedded Unit of Analysis 1
Embedded Unit of Analysis 2
TYPE 2
TYPE 4

(Source: Yin, 2018)

e. Case studies in Information Systems

Case studies are widely used in information systems research (Benbasat et al., 1987; Myers & Avison, 2002; Orlikowski & Baroudi, 1991). Benbasat, Goldstein, & Mead see case studies as the most appropriate method for researching information systems, considering that the field has changed from a technical into a more managerial and organisational one. There is also a need to consider the interrelations between innovation, regulations, contexts, and industry specificity. Viewing the method as “well-suited to capturing the knowledge of practitioners and developing theories from it” (ibidem, p. 300), they identify three reasons for adopting it. Theories can be generated from practice, questions related to reasons and modalities for complex contemporary processes can find an answer, and under-explored and innovative fields can be investigated.

3.6 This research

The following part presents an overview of the research. It first illustrates the research approach chosen considering research questions, aims and objectives. It then presents the details of the data collection, how it evolved, and how data was analysed.

3.6.1 Research approach

To answer the research question (*How can DP and the FAT principles be applied by organisations during the introduction and use of AI systems and in their digital innovation strategies?)* the researcher considered different data collection methods. The use of interviews was initially deemed sufficient to explore people's experiences. However, it soon became obvious that such an approach was not adequate. Many differences amongst sectors were being identified, and people were acting in contexts impacted by several factors. Focusing only on interviewing people was not enough. Thus, to answer the research question, the most appropriate methods were considered interviews and case studies. They were viewed as the right instruments for exploring organisational praxis, and the experiences, perceptions, and understanding of various people.

3.6.2 Research approach and research question, aims and objectives

In order to better understand the connection between the research approach and research question, aims and objectives, the table below (3-3) offers an overview of the three parts.

Table 3-3 Overview of the research approach and interview questions

RESEARCH QUESTION			
How can DP and the FAT principles be applied by organisations during the introduction and use of AI systems and in their digital innovation strategies?			
AIMS			
1.Understand the relationship between AI and DP and how they can inform each other in the context of legislation and digital innovation	2. Examine the extent to which individuals who are introducing/using AI and DP roles understand AI, DP, and FAT principles	3. Understand the impact of DP on organisations that are introducing/using AI, and vice versa	4. Produce guidance for organisations to support the application of FAT principles in their AI&DP Management

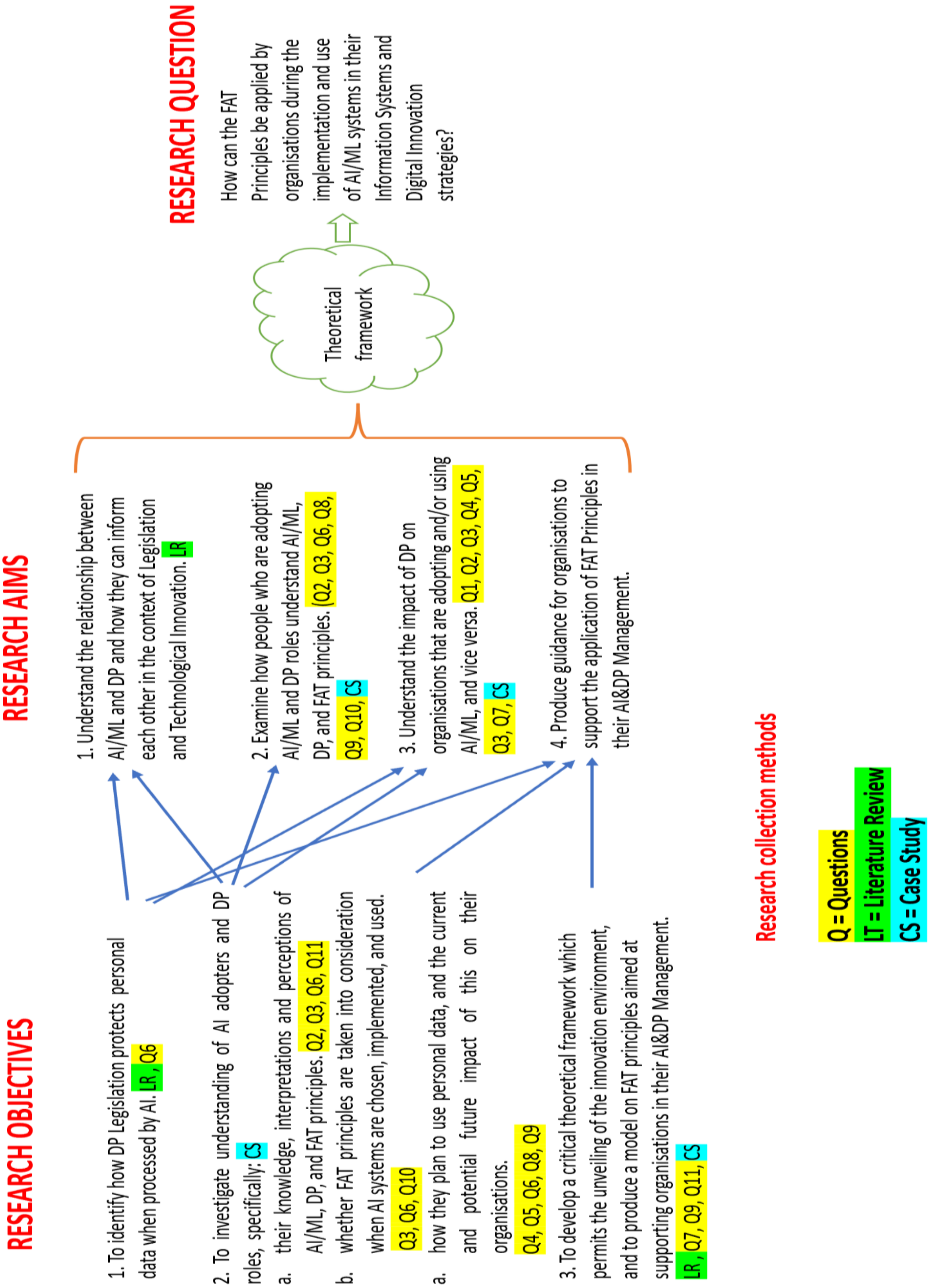
OBJECTIVES		
1. To identify how DP legislation protects personal data when processed by AI	2. To investigate the level of understanding amongst AI adopters and users and DP roles, specifically: a. their knowledge, interpretations and perceptions of AI, DP, and FAT principles b. whether the FAT principles are taken into consideration when AI systems are chosen, implemented, and used c. how they use personal data, how they plan to use it, and the current and potential future impact of this on their organisations	3. To develop a critical theoretical framework that permits the unveiling of the innovation environment, and to produce a model on FAT principles aimed at supporting organisations in their AI&DP Management

Such a table helps introduce the figure (3-3), which illustrates the relationship between the three parts, and the specific method used to gather information. For example:

- The literature review was used to evaluate the state of the art with respect to AI, DP, FAT and AI management, and how those were connected. The review helped identify gaps and comprehend how personal data processed using AI is protected under the current DP legislation.
- Interviews were used to comprehend the experience and understanding of SMEs.
- Case studies - inclusive of interviews, document analysis and observations - were used to identify organisational practices, showing up common challenges, and the experience and understanding of people in their organisational contexts.

The use of multiple sources permitted the identification of the complexity of internal phenomena.

Figure 3-3 Research collection methods and aims, objectives, and research question



RESEARCH OBJECTIVES

1. To identify how DP Legislation protects personal data when processed by AI. LR, Q6

2. To investigate understanding of AI adopters and DP roles, specifically: CS

a. their knowledge, interpretations and perceptions of AI/ML, DP, and FAT principles. Q2, Q3, Q6, Q11

b. whether FAT principles are taken into consideration when AI systems are chosen, implemented, and used. Q3, Q6, Q10

a. how they plan to use personal data, and the current and potential future impact of this on their organisations. Q4, Q5, Q6, Q8, Q9

3. To develop a critical theoretical framework which permits the unveiling of the innovation environment, and to produce a model on FAT principles aimed at supporting organisation in their AI&DP Management. LR, Q7, Q9, Q11, CS

RESEARCH AIMS

1. Understand the relationship between AI/ML and DP and how they can inform each other in the context of Legislation and Technological Innovation. LR

2. Examine how people who are adopting AI/ML and DP roles understand AI/ML, DP, and FAT principles. (Q2, Q3, Q6, Q8, Q9, Q10, CS

3. Understand the impact of DP on organisations that are adopting and/or using AI/ML, and vice versa. Q1, Q2, Q3, Q4, Q5, Q3, Q7, CS

4. Produce guidance for organisation to support the application of FAT Principles in their AI&DP Management.

RESEARCH QUESTION

How can the FAT Principles be applied by organisations during the implementation and use of AI/ML systems in their Information Systems and Digital Innovation strategies?

Research collection methods

Q = Questions
LT = Literature Review
CS = Case Study

Source: Chiara Addis

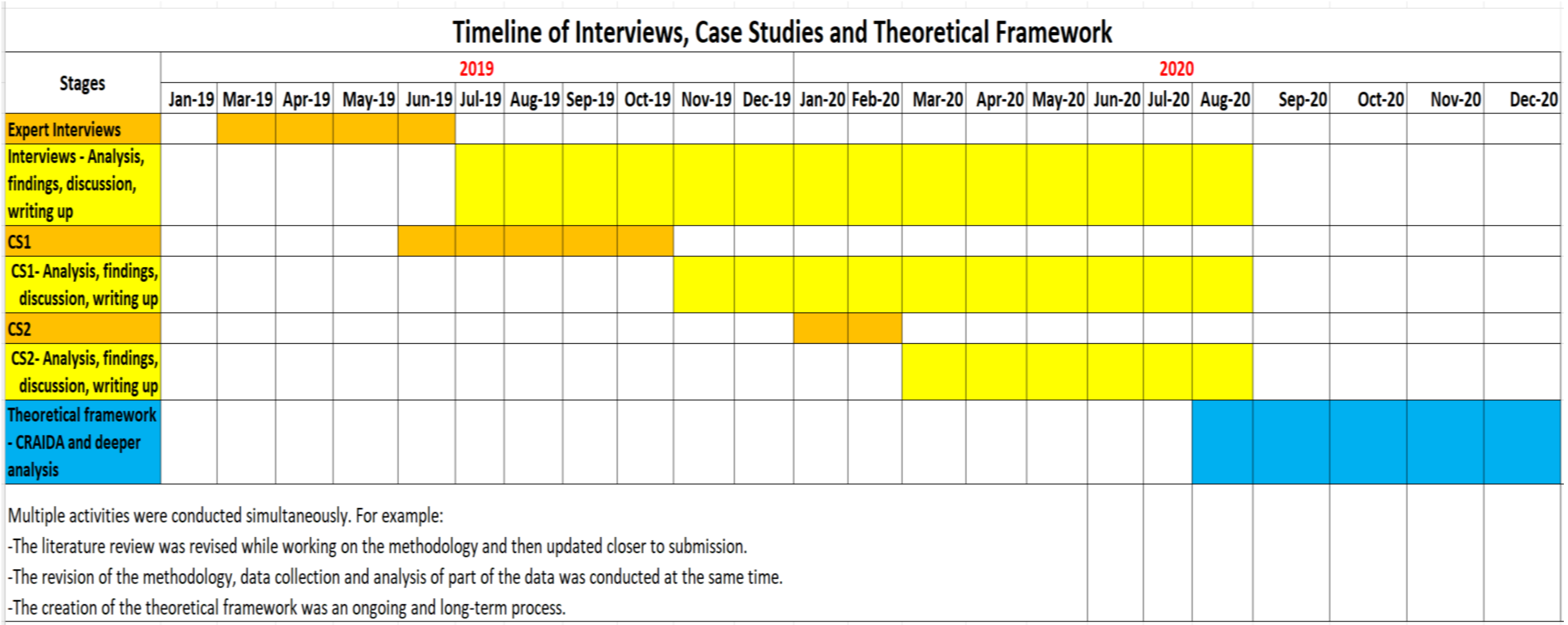
In the following, the details of each data collection method are presented.

3.6.3 Data collection methods

As seen in 3.5, to answer the research question, the researcher selected interviews and case studies as data collection methods. The interviews were conducted with a group of SMEs. They were interviewed prior to the case studies, providing crucial information that shaped the two case studies conducted immediately afterwards. An explanation of the extent to which the expert interviews informed the cases is provided at the end of the next section.

The table below presents the schedule for the expert interviews, case studies, and theoretical framework. A more elaborated timeline of the research is included in Appendix C.

Table 2-4 Timeline of data collection, analysis, and theoretical framework



Timeline of Interviews, Case Studies and Theoretical Framework

Stages	2019											2020											
	Jan-19	Mar-19	Apr-19	May-19	Jun-19	Jul-19	Aug-19	Sep-19	Oct-19	Nov-19	Dec-19	Jan-20	Feb-20	Mar-20	Apr-20	May-20	Jun-20	Jul-20	Aug-20	Sep-20	Oct-20	Nov-20	Dec-20
Expert Interviews		🟧	🟧	🟧	🟧																		
Interviews - Analysis, findings, discussion, writing up						🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨				
CS1					🟧	🟧	🟧	🟧	🟧														
CS1- Analysis, findings, discussion, writing up										🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨	🟨				
CS2												🟧	🟧										
CS2- Analysis, findings, discussion, writing up														🟨	🟨	🟨	🟨	🟨	🟨				
Theoretical framework - CRAIDA and deeper analysis																			🟦	🟦	🟦	🟦	🟦

Multiple activities were conducted simultaneously. For example:
-The literature review was revised while working on the methodology and then updated closer to submission.
-The revision of the methodology, data collection and analysis of part of the data was conducted at the same time.
-The creation of the theoretical framework was an ongoing and long-term process.

(The complete timeline inclusive of all research activities can be found in Appendix C)

a. Expert insights survey (E)

Expert interviews were chosen as a starting point for the project, because the opinions and experiences of people who have expertise in AI, ML and DP and Business Technology and who have worked with different organisations in different sectors were assumed to being able to provide insights into the implementation and use of AI drawing upon their past and current assignments. The selection criteria included the consideration of professional roles and positions within companies or projects, years of expertise, and prominence and visibility within professional networks.

Participants were identified via professional and personal contacts (e.g., former managers and colleagues working for digital and technology organisations), and extended professional networks (British Computing Society, LinkedIn, Institute of Directors).

Being aware that organisations and individuals are less willing to participate and cooperate if they perceive potential negative implications, the researcher paid particular attention to the planning and execution of the interviews. Due to the intense debate on AI, DP, and ethics occurring at the time of the research, special care was given to providing assurance on data confidentiality and participant anonymity.

Data was collected via semi-structured open-ended interviews which permitted interviewees to answer with detailed responses, allowed more flexibility than other types of interviews, and gave the possibility to adapt the questions according to different sectors.

65 potential participants were carefully selected amongst SMEs who were particularly active and well known within their respective domains (AI, DP, and business technology).

Nine participants agreed to be interviewed: four experts in ML, two experts in technology and ethics (two senior managers), one CTO, one DP consultant and one privacy lawyer. Participants worked across a range of sectors and for private and public organisations.

They all had considerable expertise in the research topics and included individuals who were part of expert groups on AI (UK Parliament) and GDPR (European Commission).

Table 3-5 Details on experts

BACKGROUND OF THE NINE EXPERTS
1. CTO, expert of digital strategy, governance, and emerging technology (advisory board member of the UK parliamentary group on AI)

2. ML researcher and former CTO. Currently working on computer vision in healthcare (i.e., predicting diseases from medical images)
3. Senior consultant. Technology leader, engineer and ML developer. They developed services and platforms, and led several digital transformation projects in the private and public sectors and across various countries
4. Technology research director, and security and ML expert
5. Lawyer practising in data protection. GDPR expert at the European Commission.
6. Data protection consultant, trainer and blogger
7. AI ethics expert and AI advisory board member for the UK Parliament
8. CEO of a health start-up and ML researcher
9. CEO of a start-up and AI ethicist

They were interviewed between March and June 2019. Six interviews were carried out face to face and three via Skype. Interviews lasted for about an hour. They were interviewed only once.

Questions were based on three main areas identified from the review of the literature:

1. AI management and strategy - Questions focused on organisations and leadership (reasons for and means of acquiring AI, level of understanding of AI, the involvement of vendors), and data and risks (data type and quality checks).
2. GDPR and processes - Questions aimed to understand what organisations had done with regard to new requirements, AI, their relationship to internal processes, and their vision of the future of AI and DP.
3. FAT - Questions explored the thinking and the application of the FAT principles, the involvement of stakeholders, and AI specific challenges and limits.

The detailed questions are presented below.

Table 3-6 List of questions for the experts

EXPERTS - INTERVIEW QUESTIONS

1. AI MANAGEMENT AND STRATEGY

1. Who is responsible for choosing AI in the organisations you work with? (e.g., CIO, or CEO, or board?) Do you know the reasons for choosing it, and which specific technology has been chosen? (Research Aim 3)

2. Do you think leaders and managers understand the technology? Is there a specific training programme for staff? (R. Aims 2, R. Aims 3, R. Objectives 2)

3. Are vendors selling the technology and working with the organisation in implementing it? (e.g., as processors). (RA2, RA3, RO1, RO2)

4. What kind of data is being used? Is there a system in place to control the quality of data (both input and output data)? (RA3, RO2)

5. Are organisations increasing their risk tolerance and risk appetite? If yes, are organisations considering any specific AI related risks? (e.g., location of learning) (RA3, RO2)

2. GDPR AND PROCESSES

6. The GDPR prescribes new requirements with regards to AI, e.g., DPIA (new requirement for private organisations), right of explanation/information, and right to human intervention. What are your experience and your opinion with regards to them? (RO1, RA2, RA3, RO2)

7. Are those responsible for AI and data protection working together for specific activities? If yes, are there specific processes in place? (RA3, RO3)

8. Do you see any difference with regards to context, both in terms of data protection compliance and AI? (e.g., sector or industry, regulated/unregulated). (RA2, RO2)

9. How do you see Data Protection and AI in the future? Will the increasing amount of personal data processed by AI be sufficiently protected? (RA2, RO2, RO3)

3. FAT PRINCIPLES

10. Discussions on fairness, accountability, and transparency (FAT principles) are growing. Have they been discussed in the organisations you work with, or between stakeholders, while implementing or applying AI? Do you think they should be discussed? (RA2, RO2)

11. Is there anything AI systems should not be doing, now and in the future? (e.g., limits in some sectors?) (RA2, RO2, RO3)

12. Is there anything else you would like to add?

Participants commented on what they had seen and experienced while working during their assignments in many organisations. Their responses provided indications of trends, information from their industries, and insights from their current assignments.

b. How the interviews informed the case studies

Conducted first, the interviews revealed crucial information for the research, design, strategy, and understanding of the case studies.

They aimed to provide an initial overview of different industries. The intent was to collect as much information as possible in order to understand what was happening in different sectors.

Questions were formulated for SMEs who had experienced those contexts, and they were asked to report their interpretation of what they had seen. Due to the desire to optimise the time with the participants, and to understand their practices in different contexts, the questions were both manifold and very detailed, inviting long and nuanced answers.

This approach worked well and provided rich insights into the work experience of experts and their understanding of the key topics explored in this research.

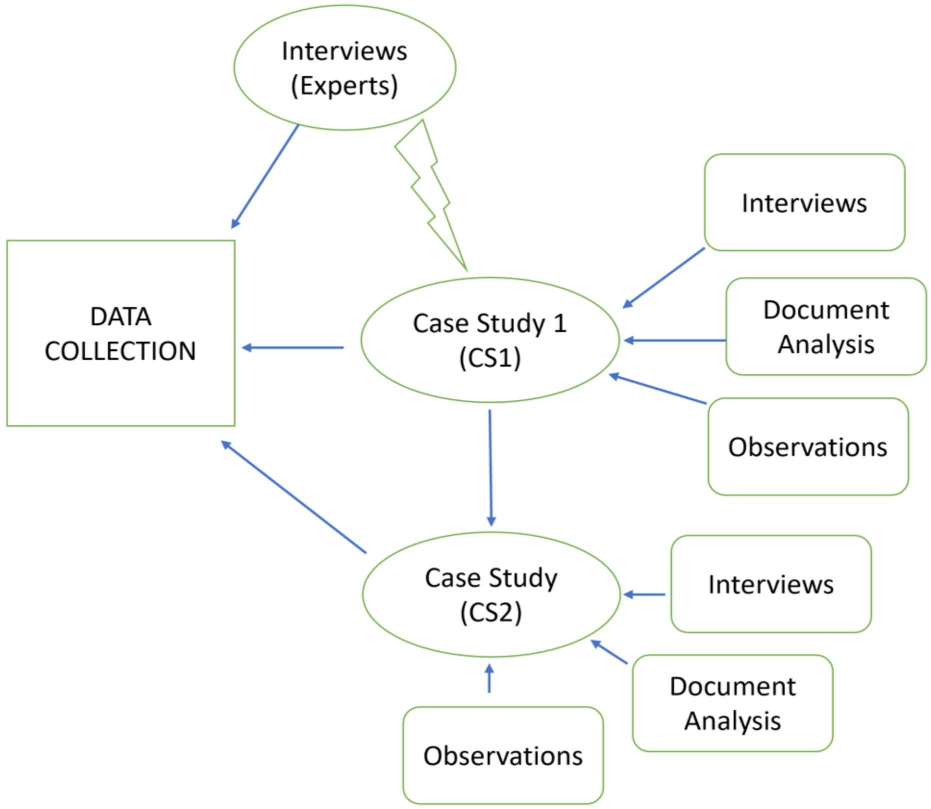
The data generated in these interviews is presented and discussed in close detail in Chapter 4. The main purpose of the discussion in this section is to show how reflections on the interviews with experts informed the design of the case studies, and also the methodological strategy adopted for them.

Importantly, experts were quite explicit in their emphasis of the existence of different practices of regulation in the private and public sector. They described the public sector as much more regulated than the private one. The realisation of a stark difference between the two sectors led to the decision to choose one case study from within the private sector and one from the public sector, allowing for an exploration of decision-making practices in different contexts of governance and regulation. Overall, this resulted in a more complex analysis of the interplay of different elements in AI and DP management. The expert interviews allowed for a good understanding of key processes and challenges of AI and DP management in different sectors. This helped to refine the interview schedule used in the case study.

Many of the questions that had worked well and had generated focused and valuable accounts were also included in the repertoire of questions for the case studies. Other questions covering areas that were expected to be less relevant in the case of the two organisations were not included. This included, for example, more generic questions about the limitations and the future of AI and questions that aimed at comparing practices in different sectors or organisations. Moreover, the expert interviews permitted the identification of some key areas to be explored in closer detail in the case studies. This included for example the realisation of the need to focus closely on the ecosystem or organisational context as a factor in decision-making, the significance of the role of various stakeholders, and detailed processes and phases within the decision-making processes.

Further conclusions were drawn with regards to the interview technique, aiming for a tighter and more focused discussion with the key actors in the case studies. Less structured questions were needed in order to give the participants the possibility to talk more freely. Thus, the questions for the case studies were fewer, more focused, and less structured. This permitted participants to be less constrained in their replies. Furthermore, a more flexible approach was adopted. While the same set of questions was put to every participant of the case study, some parts were more appropriate for leaders (e.g., around strategic decisions made by the board), and others were more appropriate for technical roles (e.g., the question around ML and the data). Providing participants with more space to talk and expand on some topics proved to be the right choice. Figure 3-4 illustrates the position of the expert interviews in the wider strategy of data collection and its impact on the case study research).

Figure 3-4 Link between the interviews with the experts and the case studies



Source: Chiara Addis

c. Case studies: case study 1 (CS1) and case study 2 (CS2)

Two case studies were conducted after the interviews with experts.

The case studies focused upon UK organisations implementing innovative AI projects. Organisations were identified via university and extended professional networks (British Computing Society, LinkedIn, Institute of Directors). 25 organisations were contacted, two selected.

-Case Study 1 (CS1)

CS1 is a UK Higher Education organisation. Established over 100 years ago, the organisation strongly focused on technology to train the local workforce. Over the years, it has developed strong partnerships with industry, local institutions, and international affiliates. Science, technology, and business are still some of its distinctive features. Different schools provide study and research activities in a broad range of disciplines, e.g., business and management, technology, health, applied science and arts. With over 30,000 students, growing diversity is a key factor. The complexity of the student body has increased in the last few years. BAME

(Black, Asian and minority ethnic) students are numerous. The organisation wants to understand better the diversity of students' needs, improve support and engagement, and prevent students from interrupting or stopping their studies. The organisation invested heavily in the use of digital technology.

The organisation was selected as a case study after considering some of its projects related to the improvement of the well-being and retention of students, namely, because of its ethical aims. ML was going to be used in one of their projects.

Nine participants were interviewed. They had various roles within the organisation, and considerable expertise in their areas. Participants were identified and chosen after considering their direct involvement in the project or their impact on it.

Table 3-7 Background CS1 participants

CS1 PARTICIPANTS' BACKGROUND
CS1/1. Student engagement manager
CS1/2. Leader. Dean of students (former leader at various higher education organisations)
CS1/3. Leader. Organisation' registrar and secretary (former registrar and leader at various higher education organisations)
CS1/4. DPO and lawyer
CS1/5. IT project manager (consultant, over 20-year experience in the private and public sectors)
CS1/6. ML developer (consultant)
CS1/7. Information governance officer (DP expert)
CS1/8. Head of student experience (former senior manager at various higher education organisations)
CS1/9. Inclusion and diversity manager (former equality and diversity manager at various public institutions)

Participants were interviewed between June and October 2019. Multiple data sources and methods were used to understand the research context:

- Semi-structured interviews - eight interviews were conducted face to face and one via Skype. Interviews lasted for about an hour.
- Document analysis - two business cases, DPIA, user guide, privacy policy, promotional material, and website.
- Observations - direct observation of social dynamics and environmental conditions was conducted during electronic communication and fieldwork, providing insights on internal dynamics and culture.

-Case Study 2 (CS2)

The second case study is a small UK start-up (DIP) that provides digital transformation and compliance services to companies in various sectors. The start-up was created by senior managers operating in the financial sector after realising the need for a more efficient compliance and identity management systems. DIP was selected for their specific combination of values and for the characteristics of the project they were developing around digital identities. DP, sustainability, and ethics are key features in their work and their desire to increase trust in business was particularly appealing for this research. The project aimed to simplify data exchange between organisations and individuals, reduce the cost of compliance, and strengthen data subjects' consent. Consent is seen a key factor in increasing transparency, control over data, and agency of subjects. The use of facial recognition and the strong emphasis on consent were crucial in choosing DIP as the second case study. The work by Edwards & Veale (2017) on lawful bases of processing and consent presented in 2.4.4 increased the desire to explore DIP's use of consent and therefore to use it as the second case study.

Table 3-8 Background CS2 participants

CS2 PARTICIPANTS' BACKGROUND
CS2/1. CEO and DPO. Leader with over 20-years' experience in management, technology, risk, compliance, and HR
CS2/2. Director. Leader experienced in digital transformation, compliance, security, and fraud prevention

They were interviewed in January and February 2020. Multiple data sources and methods were used in this case as well.

- Semi-structured interviews - two interviews were conducted face to face. The interviews lasted for about an hour.
- Document analysis – two DPIAs, privacy policy, promotional material, and website.
- Observations - direct observation (electronic communication and fieldwork).

All participants (CS1, CS2) were interviewed only once. The interview questions in both case studies were still based on three main areas identified from the review of the literature (AI. GDPR, FAT). However, as explained in point b, the questionnaire was updated after the interviews with the experts.

Table 3-9 List of questions for the case studies

CASE STUDY - INTERVIEW QUESTIONS

1. AI MANAGEMENT AND STRATEGY

1. Why is your organisation buying/developing/using AI? Are vendors involved?

2. What kind of AI and data are used? ML? Any specific checks? Any specific risk?

3. Do you think people (management and staff) understand the technology?

2. GDPR AND PROCESSES

4. Are you aware of the GDPR changes? How did your organisation implement them? Any specific requirement for your sector?

5. Are the roles in charge of AI and data protection working together? How?

6. How do you see data protection and AI in the future? Does your organisation want to expand AI and use more data in the future?

3. FAT PRINCIPLES

7. Have fairness, accountability, and transparency been discussed in relation to this project, the GDPR, stakeholders or in general? Any specific issue?

8. Do you think they should be discussed, and how?

9. Is there anything else you would like to add?

The two methods generated a diverse and rich amount of data. This was then analysed following a specific approach.

3.6.4 Analysis

The analysis of the data was conducted using a thematic analysis approach (Braun & Clarke, 2006, 2017; Lester et al., 2020; Vaismoradi et al., 2016). Such an approach permitted the identification, analysis, and reporting of patterns (themes) in the data (Braun & Clarke, 2006). Thematic analysis provides a systematic procedure to interpret complex datasets through the generation of codes. In acts of careful and repeated reading of interview transcripts and documents, thematic analysis uses coding through the creation of thematic categories to identify meanings that are prevalent within the dataset. The use of codes facilitates a structured and systematic (comparative) reading across different sources that allows researchers to explain certain phenomena in a more complex manner. In the first instance, coding provides an overview and a substantial knowledge of the content of data. On that level, coding is descriptive and helps to organise the data. Beyond that, working through the material gathered under the rubric of certain codes allows to uncover the prevalence of – and nuances within – experiences, presentations or opinions expressed (Braun & Clarke, 2006, 2017). Coding thereby aids an interpretive analysis by identifying themes – and the connections among them – which are worthy of further discussion.

In this thesis, the overall set of emerging codes was used to make strategic decisions about which aspects of the data to explore and present within more detailed accounts.

Data included the transcripts from the interviews, documents provided by and on the organisations, and the notes from the observations. In the context of this research, thematic analysis involved the following steps:

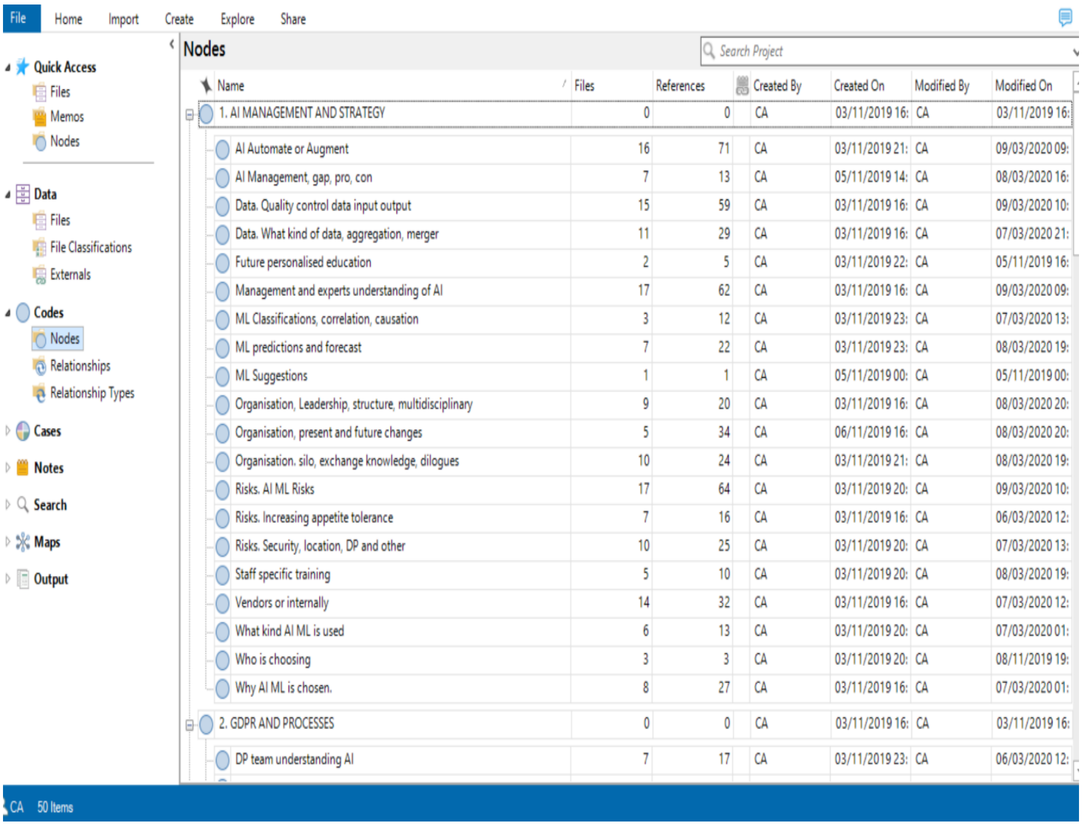
1. Data was gathered, prepared, managed, and organised (e.g., specific folders were created for the recordings, and notes from the observations were converted into electronic documents) to facilitate careful analysis.
2. The interviews were transcribed (manually – verbatim transcription).

3. The first level of analysis involved careful reading of the transcript for familiarisation of the data and some initial codes were generated by writing down some preliminary ideas.

4. While coding is usually performed within an inductive analysis “without trying to fit it into a pre-existing coding frame” (Braun and Clarke, 2006, p. 83), such a process does not happen in an epistemological vacuum (ibidem). The nature of the research question rendered it appropriate to orient coding towards an exploration of the overarching key themes of the project, i.e., AI, GDPR, FAT. Starting out with this threefold perspective, ultimately six main nodes were generated. These included: 1) AI management and strategy; 2) GDPR and processes; 3) FAT principles; 4) Future AI/ML and DP; 5) Power and stakeholders; 6) Post-project outcomes.

5. The act of coding, i.e., assigning significance to the data via “a short, descriptive word or phrase” (Lester et al., 2020, p. 100), was performed using the software NVivo (Figure 3-6). The transcripts were uploaded onto NVivo. During the analysis of a particular interview, portions of the texts were allocated to specific nodes leading to the list of six main codes mentioned above, and various sub-nodes, in the course of which patterns emerged, allowing for the definition of key themes and categories. Coding thereby involved a networked and hierarchical organisation of categories within each respective thematic context. An example for a nodal cluster can be seen in Figure 3-5 below.

Figure 3-5 An example of coding with NVivo



Name	Files	References	Created By	Created On	Modified By	Modified On
1. AI MANAGEMENT AND STRATEGY	0	0	CA	03/11/2019 16:	CA	03/11/2019 16:
AI Automate or Augment	16	71	CA	03/11/2019 21:	CA	09/03/2020 09:
AI Management, gap, pro, con	7	13	CA	05/11/2019 14:	CA	08/03/2020 16:
Data. Quality control data input output	15	59	CA	03/11/2019 16:	CA	09/03/2020 10:
Data. What kind of data, aggregation, merger	11	29	CA	03/11/2019 16:	CA	07/03/2020 21:
Future personalised education	2	5	CA	03/11/2019 22:	CA	05/11/2019 16:
Management and experts understanding of AI	17	62	CA	03/11/2019 16:	CA	09/03/2020 09:
ML Classifications, correlation, causation	3	12	CA	03/11/2019 23:	CA	07/03/2020 13:
ML predictions and forecast	7	22	CA	03/11/2019 23:	CA	08/03/2020 19:
ML Suggestions	1	1	CA	05/11/2019 00:	CA	05/11/2019 00:
Organisation, Leadership, structure, multidisciplinary	9	20	CA	03/11/2019 16:	CA	08/03/2020 20:
Organisation, present and future changes	5	34	CA	06/11/2019 16:	CA	08/03/2020 20:
Organisation. silo, exchange knowledge, dilogues	10	24	CA	03/11/2019 21:	CA	08/03/2020 19:
Risks. AI ML Risks	17	64	CA	03/11/2019 20:	CA	09/03/2020 10:
Risks. Increasing appetite tolerance	7	16	CA	03/11/2019 16:	CA	06/03/2020 12:
Risks. Security, location, DP and other	10	25	CA	03/11/2019 20:	CA	07/03/2020 13:
Staff specific training	5	10	CA	03/11/2019 20:	CA	08/03/2020 19:
Vendors or internally	14	32	CA	03/11/2019 16:	CA	07/03/2020 12:
What kind AI ML is used	6	13	CA	03/11/2019 20:	CA	07/03/2020 01:
Who is choosing	3	3	CA	03/11/2019 20:	CA	08/11/2019 19:
Why AI ML is chosen.	8	27	CA	03/11/2019 16:	CA	07/03/2020 01:
2. GDPR AND PROCESSES	0	0	CA	03/11/2019 16:	CA	03/11/2019 16:
DP team understanding AI	7	17	CA	03/11/2019 23:	CA	06/03/2020 12:

(Source: Chiara Addis)

6. Hierarchical coding was used when specific and narrow classifications were identified. For instance, further layers of coding were elaborated within the code quality control of data (e.g., internal/external checks and remediation offered to end-users). The researcher engaged inductively with the data through coding and later identified categories (by linking codes). Recurrent or significant themes (e.g., role of stakeholders, AI-human interaction) helping to answer the research question were then identified. Themes that appeared repeatedly and revealed significant connections or allowed for consistent explanations were then chosen for detailed discussion.

7. The interpretation brought themes and explanations from within different nodal contexts in conversation with each other and debated theoretical perspectives salient within the discipline.

8. The report with the analysis was written up, including both conceptual analysis and extracts from the data linked to the themes to address the research question, aims, and objectives. It also fed into the development of the theoretical framework.

The use of multiple data sources and methods helped to triangulate and understand the research setting and contexts. While the two cases varied considerably due to size and sector, some elements identified in these respective cases (e.g., stakeholder management) can certainly offer precious information for similar situations and contexts. Thematic analysis provided a detailed understanding of organisational practices, individual experiences, and the implications of the implementation of AI for DP matters.

3.7 Conclusion

This chapter has presented the research methodology deployed to understand the praxis of AI/ML, DP, and FAT principles. It reflected upon different research philosophies (positivism, interpretivism, and critical realism), modes of research reasonings (deductive, inductive, abductive), and methodological approach (quantitative, qualitative, mixed methods).

The chapter then explained the reasons for choosing interpretivism, inductive and qualitative approaches, presented the details of the chosen collection methods - interviews and case studies - and how data was analysed. A qualitative multi-method approach was thought to be most adequate for the complexity of the research questions.

The results of the data collection allowed for a comprehensive understanding and analysis of organisational practices and experiences that will be presented in the following chapters.

The analysis of the experts' opinions and experiences will be presented in Chapter 4, while the analysis of CS1 and CS2 will feature in Chapters 5 and 6. Chapter 7 will illustrate the theoretical framework of the research, and how the framework has been used to interpret the data.

CHAPTER 4: EXPERT INSIGHTS SURVEY

4.1 Introduction

This chapter presents the interviews with a group of SME experts (E) in AI, DP, and business technology. The survey aimed at understanding the dominant trends on AI adoption, GDPR implementation, and FAT understanding in UK organisations and the analysis is discussed considering the three main research areas, AI, GDPR and FAT. After presenting the discussion on AI management that focuses on the main actors and risks, the chapter then reveals details on GDPR implementation and information on how FAT principles are understood and operationalised within organisations. The chapter then ends with a comprehensive discussion of the key elements emerging from the analysis of the data.

4.2 Group of experts overview

This part of the research comprises interviews with experts, inclusive of three technology business technology experts (E1, E7, E9), 4 AI/ML experts (E2, E3, E4, E8), one privacy lawyer (E5), and one DP expert (E6). The participants had considerable expertise gained in different sectors and industries and their interviews provided information on trends and insights into the praxes of various organisations (further information in 3.5).

4.3 Analysis

In this section, the key themes identified in the analysis of the interviews are presented, followed by a discussion considering the overall lessons that can be learned and how they inform the rest of the research.

4.3.1 AI management

a. Strategic decisions and stakeholder management

The participants reported that organisations acquired AI technologies for different reasons. There is pressure originating from boards, senior management or innovation areas within the business that is driven by the desire to improve operational efficiency, reduce costs, and maintain competitiveness. In some cases, this pressure arises from worries about the sustainability of the organisation. However, implementing AI in more traditional organisations can be challenging (E1, E7): *'[Organisations] that have been around for more* *than eight years, they have a problem […] an industrial core, how do they embed AI into their business? How do they change their companies?'* (E1). At the same time, individuals and teams were reported becoming informed about AI, seeing the potential for improving their activities, and suggesting it to management. Organisations were becoming more aware of the data they hold, and its potential value. They suddenly realise they are sitting on lots of data, and that could do something. For instance, '*getting more insights using machine learning and that might be useful for the business, or they may create systems that could be monetised'* (E1). The availability of open-source technology and cloud-based frameworks makes AI easily and quickly accessible. Many start-ups were reported taking more risks to react quickly to market changes (E4, E5, E8). This supports the literature on the pace of digital innovation, and its impact on risks and compliance. Many big organisations were developing their own systems, often without central governance oversight of what was being done in different business areas. This is intensified in organisations with strong silo mentalities and many different computer systems (E1, E9). The need for a more holistic approach was considered necessary to mitigate the risks, but not many organisations were reported to be adopting such a strategy (E9).

Executives and management were reported to be displaying a generally low knowledge of AI, but this was slowly improving (E8). They often acquire AI without having the capacity to develop, implement or sustain it (E4, E8). Leaders were signing off systems they did not properly understand (E4), including the implications for security and DP compliance:

> *They themselves don't fully understand what they are doing, and they probably have a very high level of abstraction. They don't know which questions they should be asking and do not understand the deeper level to be worried about, like biased algorithms* (E4)

The low level of specific knowledge often translated into uncertain strategies and high expectations. Leaders were facing big decisions. Many were choosing AI automated models and were reducing the reliance on the workforce (E1), while some others (mainly in the public sector) were concerned about the impact on employees (E3).

Vendors were reported to be providing off the shelf technology when organizations need tailored solutions that generic solutions cannot provide. Some vendors are often more aware of security risks and obligatory compliance and were informing clients. Many were also using clients' data for their innovation, due to contracts that allow data capture, which confirms their role as joint controllers/controllers and not mere processors. Those more aware were reported to be very careful in gathering data and storing it (E4).

b. Different risks and their perception

Risk awareness and risk management were mentioned as crucial factors in GDPR compliance and AI management. GDPR was identified as having increased the awareness of risks. As the Regulation allows a risk-based approach, managing the risks strategically seemed to be a common strategy for organisations. Some risks were being reduced with more confidence in areas with fewer ambiguities, e.g., data storage (E4), and increased in others (e.g., data aggregation). There were still uncertainties, for example, around the withdrawal of consent from data being used to train ML models (E4, E5).

Large and high-profile organisations generally had a low-risk appetite. For instance, they were reported adopting a cautious approach before starting new initiatives (E5), taking advice on many aspects (e.g., security, data location and data access), and taking their time to understand how the GDPR changes the way they operate (E7).

Many start-ups were believed to be less aware of the GDPR requirements and to have a higher-risk appetite. They were reported to be taking more risks and focusing on the quickest way to get products to market (E4, E5, E8), often without understanding the implications. The growing practice of using open-source technology is helping to develop the technology, often with other consequences: '*The whole ethics around it, transparency around it, I think people see it more as a barrier and they are reluctant to engage in that debate because that might stop them from getting the benefit*' (E4).

Having a smaller number of data scientists was also considered to play an important role in risk awareness within small companies and start-ups, something these organisations did not seem to fully understand.

Vendors having AI and GDPR competencies were said to be investing in research to reduce risks for themselves and their customers. This was illustrated by E4 whose company was looking at a number of methods, such as homomorphic encryption and differential privacy.

Particularly interesting were the specific ML risks identified by the participants. These included discussion of the following themes:

- Data source. The source of training data can carry some high risks, for example, the collection of uncontrolled data from the Internet (E2) and the acquisition of data from other countries with different legislation (e.g., China). A trade of health data was also reported: '*There has been a lot of concern [...] So, using this data for personalised marketing, or potentially even selling this data to insurance companies. People will be denied healthcare, you know, in countries with private healthcare*' (E8).

- Training, deployment, and loss of accuracy. A gradual reduction in the efficiency in some ML algorithms was reported, with systems developed in the lab with training data showing a high accuracy rate, which then drops when deployed within the real world: '*Some systems might be able to retrain it [...] others would continue to degrade performance and accuracy because they are not retrained on new real data'* (E4).

  Some research reported the development of tools to detect a drop in efficiency and the need to retrain on new samples (E4), something that the out-of-the-box systems were said to be lacking. Not many organisations were believed to be aware of this reduction of efficiency, which was compared by E4 to old Intrusion Detection Systems, good for detecting only known threats.

- Reproducing the past. Societies change constantly, and this impermanence should be considered in order to update the data at intervals (E2). ML systems that do not adapt '*are just looking to match the things we were aware of in the past [...] this is a serious problem for senior people if they are not educated, as the name AI makes you think that that is already happening'* (E4). The opposite is also possible when ML adapts to new data sets and '*forgets*' about what was happening in the past, becoming less accurate (E8).

- Predictions and confidence in the model. Understanding the correct prediction made by ML in healthcare can be challenging:

  ...you don't know if the machine randomly got it wrong. You can look at how does the system behave on average...but for the individual, you don't know whether or not it's got it right or wrong, some models can tell you about the confidence in their model (E2).

- Data manipulation, reverse engineering, or adversarial data. These were reported in general as not being considered enough by organisations, with some differences due to data sensitivity and contexts.
- Automated and Augmented AI models. Several participants raised the issue of potential risks arising from automated systems, for example, a potential lack of control, intelligibility, and accountability. Augmented AI was in general assumed to be less risky, as having a human in the loop was seen as an important factor in eliminating or lowering the risks. The human should have the '*possibility to intervene and change the decision'* (E1), or to increase the accuracy: *'If the algorithm provides 60% of accuracy, and that was fine according to the implications in that case…for 40% of uncertainty I can put a person'* (E3). Humans were reported to understand and correct mistakes in different moments of the process (E1, E3). A growing number of organisations were said to be choosing ML to support human decisions rather than replacing them (E7).
- Data access in a controlled environment. Accessing data held in controlled environments was raised as a specific DP risk:

  To train ML I need real data in a very controlled environment, with very limited access. When developers need access to that data, usually a specific environment is created for that purpose…for less mature organisations that is a real eye-opener (E2).

  After using it for training, developers are supposed to discharge the environment: '*We may request the same data next time…and this can be challenging because we don't have the same training set…to see how they both behave'* (E3). In similar situations, the number of staff looking at that data (and potentially sensitive data) can increase, e.g., when the system goes down or when the data is used to improve the service provided (E2).

Other identified risks were linked to the management of AI. Risks were frequently underestimated due to incorrect assumptions and expectations, with necessary checks and controls being ignored, or to a misunderstanding of responsibilities and accountability, for instance, by assuming that others were dealing with DP and security measures, or that all checks were done in the cloud or by the vendors implementing AI, or that other departments owned the data and the responsibility (E1).

Therefore, while some organisations were cautious or aware of taking risks, others appeared to be making the wrong assumptions or lacking the necessary knowledge or specific information to understand the DP risks associated with the use of AI.

c. Summary

Under AI management, some key elements that will inform the *RAIDIS* model in Chapter 8 were identified. For example,

- Strategic decision making, market pressure and organisational dynamics.
- Technology, instability of ML and the importance of the data source.
- People, and the impact of assumptions on processes.

4.3.2 GDPR

a. Compliance with GDPR

There was a consistent view amongst participants that many organisations were not yet GDPR compliant. Many were reported to have done only the minimum to become compliant or to not having started at all, possibly as they were waiting for their competitors to have major data breaches (E1). Others were adopting a cautious approach and taking the time to understand how to implement changes and processes that are GDPR compliant (E7). Many were said to be avoiding the debate around AI and GDPR due to its complexity and worries that this could impact on their innovation (E4).

Differences appeared to exist across sectors, and according to the data maturity and size of organisations. Large organisations invested resources in compliance (E7) and were looking strategically at GDPR (E6). This was also observed within smaller organisations with a mature data culture (E6). Organisations in the regulated sector were reported to be more mature, confirming the gap that emerged in pre-GDPR research examining organisational preparation for the Regulation (Addis & Kutar, 2018). Finance and large technology sectors were in a more mature stage in relation to awareness and the application of good practices, such as the creation of working groups '*to make sure everybody is connecting on the same page*' (E4). Organisations in the public sector were also generally more compliant, due to the sector being more regulated, and to some of the new GDPR requirements such as the DPIA having previously existed as requirements.

A lower degree of AI awareness and GDPR compliance was reported in the private sector, particularly within medium and smaller companies and start-ups (E4, E8) where DP was not seen as an issue or not even on their radar (E5). And yet, AI awareness was not necessarily higher amongst those who display DP expertise. Particularly striking was the observation made by the participant who sat on an EU Group of Experts on GDPR, where AI was reported to be largely absent amongst the topics.

b. GDPR requirements and processes

The organisations using AI to process data were not believed to be carefully thinking about their lawful bases. When implemented carefully, GDPR was also seen as a means to protect vendors using AI to develop their products, providing they could demonstrate that the balance between their and individuals' interest was carefully considered (E5).

Another significant point was related to lawful bases of processing in automated decision-making. Relying only on consent can be risky for organisations. Asking for individuals ' consent can be difficult prior to processing, and risky post-processing as data subjects can withdraw consent (E5). Other lawful bases were considered more appropriate, for instance, contract or legitimate interest, but the processing ought to be the only necessary means to achieve that purpose, sometimes difficult to prove. E5 raised an interesting point, noticing how the 'necessity' was often dependent on the business model and the understanding of the concept:

> *If your business model works by high volume decision making based on algorithms, does this in itself mean that the decision making is necessary? On a strict view, the answer is no. On a more kind of open view, the answer is maybe, maybe it is, depending on where you give other safeguards within your process* (E5)

Considering the growing interest in digital transformation, this point is particularly relevant at the current time. More organisations are choosing digital strategies which can modify or change their business models towards a more digital core, and progressive developments towards more automated business process models are expected. Will this always create a 'necessity of processing' for digital businesses? Will this make the legitimate interest a default legal justification, exonerating organisations from looking for another legal basis?

c. GDPR requirements vs praxis

The GDPR requirements regarding AI were not seen as easily achievable at this early stage of the use of the technology (E1, E3, E4). The internal organisation, sector and maturity of the organisation were all key factors in understanding and satisfying GDPR requirements.

Internal cooperation and exchange between teams dealing with IT and data, and those dealing with DP (such as DPOs and Information Governance managers) were considered particularly important. This was more common in the public sector, likely due to the fact organisations were already performing DPIAs. In general, not many organisations were reported to be performing them because they were seen as '*a luxury'* (E5), 'a *philosophy…very difficult to tie up*' (E6), and not as a valuable tool: '*I do not think that organisations necessarily recognise how useful DPIAs can be'* (E5). It was not yet clear if organisations starting to deploy or use AI were performing DPIAs or if vendors selling the technology were performing them, something considered particularly desirable and important by E5.

In contrast, some organisations in the public sector were reported performing too many DPIAs (E6), even for low-risk situations (not a GDPR requirement). While this was not seen as negative per se, it was noted that the resources used for DPIAs could be employed for other activities.

Privacy by design (PbD) was mainly considered for specific issues, and not treated as a preventive and ongoing activity (E5). The increased role of the DPO is another key GDPR requirement whose impact varies across sectors. Some organisations were reported to have concerns about the DPOs' power. This was mainly occurring in the public sector, and less so in the private sector: '*They know the DPO is an adviser, and they think of them as a lawyer and often appoint a lawyer. Those who have appointed a DPO understand it'* (E6). And yet, having a DPO does not necessarily guarantee compliance as '*years of struggle in getting themselves [DPOs] consulted*' (E6) were foreseen for DPOs. When companies in the private sector take the role of the DPO very seriously, those are usually large companies with a low appetite for risk.

d. GDPR vs strategic decisions

AI is generally acquired for a specific purpose, usually to increase efficiency, and in many cases is implemented without considering or understanding the implication for DP. Often acquired for one purpose, AI is then also used for other purposes, which could result in unlawful and unfair outcomes:

> *Technology always comes first, it's quick and easy to use. This system will allow you to do something simple, usually more efficiently. Once the system is there, they start to see patterns, and the uses start to present themselves. In the GDPR terms that is the other way round* (E6)

The purpose of processing was regarded as the preferred starting point of the process which leads to the acquisition of AI, as per GDPR. It starts from the need of the organisation and moves to the identification, acquisition, and implementation of the technology. However, this is not what was said to be happening in many cases. The technology appears to be the driver, and not the business case. Using the technology for additional purposes has massive implications for DP (E6). For instance, one problem relates to monitoring employees: *'Once the technology is there the uses occur to people…so they don't even go into it with the intention of monitoring… ' (*E6*).* The monitoring of both resources and people within organisations is increasing. For example, tracking business vehicles or employee access to premises is often done using biometrics such as fingerprints (E6). This is done for pragmatic reasons. These solutions were considered easy and reasonable when individuals are given alternative options such as entry codes to buildings. E6 envisaged a gradual increase of monitoring, inclusive of people and their performance. However, the same participant noted that decisions based on data still tend to be made by people, and not by machines via automated decision processes.

e. Summary

Under GDPR, some key elements were identified. For example,

- Strategic decision making.
- Impact of stakeholders on compliance.
- Technology, data, and processes in the full innovation cycle.

Those elements will inform the model in Chapter 8.

4.3.3 FAT

a. Accountability and fairness

The understanding of accountability varied. It was often considered as one of the easiest GDPR requirements to satisfy when linked to security, i.e., data location, storage, and access (E5). In other circumstances, its meaning was less understood, and in the case of small organisations, it was said to be '*not even on their radar'* (E5).

Accountability was reported to be particularly challenging in the case of black-box algorithms, ML systems continuously learning without any human oversight or having too high a degree of autonomy (E3, E4). However, other reasons less dependent on algorithmic intelligibility, for example, people's competencies, power, and team interactions were also reported to be impacting on compliance. E3 recalled the case of a manager who, lacking specific competencies on AI, was delegating other people in the team, but carried on owning the responsibility. *'There is a wider debate on how much responsibility bosses have…if something goes wrong, how responsible and accountable they are? '*(E3). Such situations were considered particularly problematic, as they were seen as a sign that some managers were signing off documents without understanding the consequences.

Similarly, the understanding of fairness could vary. In general, not many were reported to be thinking about fairness and AI, as *'it is a rather niche area within AI [...] the model or the solution would solve the task in the best possible way. If there are biases in the datasets [...] usually, I would say this is mainly ignored'* (E8).

Fairness was reported to be dependent on biases in the data, but also on the management's decisions on data, generally made by one person or a group of people within the organisation. Potential biases in ML system were mentioned regarding various situations (E7, E8, E9), such as data used by the NHS, institutions, or justice courts (E5). Describing a complex picture, with danger resulting from both input data (which may reflect systemic bias and social structures of discrimination) and the way machines are trained, E5 could see how biases can be present without the awareness or knowledge of the organisations. However, biases can also provide organisations with some useful excuses, as it is easier to blame AI for internal inefficiencies (E9). The implementation can be challenging when involving neural networks: '*This can provide a convenient excuse for people around things like transparency,* *when you ask the question how do you know what you are doing? [...] and the defence we see coming back a lot is, well we don't, that is the nature of the deep neuro network'* (E4).

Greater awareness around the identification and prevention of biases was reported in the last few years. However, it was not clear if, and how, this was being translated into practice. Ongoing tests to identify discrepancies in the model were seen as an expression of rigour around training, which can also facilitate the identification of discriminatory outcomes for specific ethnic groups, even when the ethnic background data is not fed into it but resulted from other factors. By looking at the data, via reverse-engineering, those factors could be identified (E3). This would constitute, and demonstrate, good rigour in face of poor explainability. However, this comes with a cost for the organisation (E3), which not many were willing to pay.

Furthermore, it was unclear where qualitative checks on data were being performed next to quantitative one, and only if an individual would bring concerns or if one of the regulators would really get involved (E5).

Fairness was also linked to management's decisions. E6 recalled how some leaders 'concerns can translate into an increased data collection. The fear of stigmatising certain categories of people, for example, by collecting data only from a specific group, or the lack of clarity on specific purposes, can lead organisations to collect data from everybody, increasing the amount of data and the related risks associated with data compliance:

> *People making these decisions are always worried [...] and they collect everything for not missing out [...] so, with data protection is easier if you have a clear purpose…always easy to justify, but it still needs to be driven by the purpose, the benefit, rather than that there is a piece of software available…*(E6)

New problems can arise when the data increases and *'you don't know what we want to know…we don't know what we're going to do…and we are going to let the data teach us'* (E6).

This was a clear call for better coordination between strategy, innovation, and data minimisation.

b. Transparency

Transparency and explainability were recognised as big issues, and Automated AI was, in general, assumed to be riskier and less transparent. A fully automated process is generally used when the benefit is on the end-user and the risk for the business seems to be low (E3). The human intervention included in Art 22.3 was already practised by some organisations (in the regulated sector).

E2 discussed the relation between intelligibility vs performance: '*what happens in practice is that you can have a model that works very well, but you don't know how it works… '*. This is especially the case with deep learning whose intelligibility surprises ML experts as well: '*DL is a massively complex decision tree, there is nothing beyond that…the amount of time I sat working on something, being frustrated, not working, and then the following day it worked… '* (E3). The implications for auditing, transparency and morality were made clear by E4 who added: '*That is the nature of deep neuro network, nobody really knows why it made that decision and why it learnt the way it learnt […] even the architect of the system doesn't know how it works or how it is making that decision '* (E4).

The GDPR risk-based approach helps organisations to minimise the risk. This lack of intelligibility was more accepted in some situations and contexts, such as screening of e-commerce reviews (E3), while in others it was seen as more problematic.

The lack of intelligibility translates into a lack of explainability. Can ML provide the organisation with a meaningful explanation?

> *The way that AI works it's completely different to the way a rational human brain sees, thinks…AI has not got a contextual understanding of what's going on. So, to explain decision-making, it's very difficult to say what AI is doing, in what counts in human terms as an explanation* (E5)

Of interest was the indication that other elements could provide the organisations with further information, similar to what the Computer-Aided Detection (CADe) can do with radiographic images. For example, the analysis of the context of processing, the purpose, and the people involved in the process were considered important elements in providing more clarity and information. The lack of explainability challenges the GDPR compliance of organisations using Automated AI systems, as autonomous decisions create the obligation to provide data subjects with a meaningful reason (Art. 22). However, even in a situation where Art 22 is not applicable (e.g., profiling without a decision), Data subjects have other instruments. Individuals can request access to their data and check whether a product is using that information: *'quite the extent to which that is appreciated, and or possible or feasible is a real and interesting tricky area'* (E5).

In general, people were reported as not being very familiar with FAT and GDPR principles, but in some more mature organisations, *'people are asking the right things in the right way [...] for example, at some high level of risk management, but they are not doing it consistently and not about the GDPR'* (E4).

Transparency, auditing, human supervision, and power to overwrite decisions made by AI, were all starting to be discussed inside organisations. However, there is little enthusiasm for engaging in this debate due to the complexity of the topics and concerns that this can become a barrier for their activities: '*They want to develop the technology and use it, engaging in that debate might stop them getting the benefits'* (E4).

c. Summary

Various key elements were identified under the FAT principles. For example, different understanding, and praxis, of FAT according to knowledge, positionality, specific technology, and external pressure. These aspects too will inform the model.

4.4 Discussion

In the following section, the key elements identified in the analysis of the data will be discussed.

4.4.1 AI management

Thanks to the GDPR and the growing adoption of AI technologies, both DP and AI are now more visible and discussed, but this does not necessarily translate into a higher understanding of them. As illustrated by the findings above, general low awareness and a limited specific knowledge amongst organisations were said to be common. This does not come as a complete surprise given that the Regulation was fairly new at the time of the interview, and the understanding of AI is still low and often unrealistic outside of expert circles. However, a higher level of awareness, compliance and collaboration between different entities was expected, given the volume of processed data, the growing adoption of the technologies and the threat of heavy fines.

It was believed that few organisations were GDPR compliant. Participants reported slowly increasing awareness in their activity, with more preventive thinking, which indicates the positive effect of GDPR on organisational awareness. Many differences were reported according to sectors, size, and maturity of organisations. Mature organisations and those operating in the regulated sector are in general more aware of the risks and of the impact the GDPR has on processes, people, and data. Some organisations were reported to be taking the time to understand its full implications and to choose effective compliance strategies. They have a good understanding of how to use privacy-enhancing tools (such as PbD) and DPIA, and how to use the GDPR strategically. This is important and indicates the organisational reach of effective approaches to GDPR, with effective compliance dependent on a nuanced understanding of its relation to the processes, people and data and ongoing consideration of DP principles and requirements. Very surprising was the reported reduced space for AI amongst those possessing high GDPR expertise at the EU level, which appears to indicate scarce awareness (at the time of the interview) of the potential risks of AI on personal data.

Small organisations and many start-ups are displaying high-risk appetite, are responding fast to the market and are less aware, and in many cases scarcely interested, in evaluating risks and GDPR compliance.

In many cases, the GDPR arguably translates into a cost that many may be unable, unwilling, or just not interested in paying. In others, the complexity or the perceived complexity of AI systems can provide organisations with the justification for avoiding deeper analysis and accepting responsibility.

Leaders are facing crucial strategic decisions, often without possessing the necessary knowledge to make informed choices that can impact on the workforce and the sustainability of the organisation. Organisations are often overestimating the suitability of IT systems, and this can become costly. They are acquiring the technology via different means, mainly developing AI systems internally or via vendors. Both involve clear risks for DP.

The relationship with vendors who are developing/implementing the technology can have direct consequences on compliance. This is often dependent on their specific knowledge, compliance, and willingness to support their customers. Those more aware of GDPR and AI were reported to be investing in research to strengthen their compliance and gaining a competitive advantage while improving the relationship with clients.

The GDPR tries to regulate the relation between the organisation and third parties, and it is quite prescriptive in doing so. However, the complexity, potential and risks of these relationships are very often misjudged by organisations that are often providing AI vendors with access to data for their innovation.

Inappropriate management of AI, and inadequate management of DP can prove very risky for organisations. Organisations are not always aware of the specific risks or do not have the organisational capacities to quickly identify or mitigate them. The insights provided by ML and Business Technology experts proved very valuable for identifying some specific risks linked to different areas, such as data, technology, system and organisational structure and organisational practices.

ML experts exposed some specific characteristics of ML which are often overlooked or ignored. The referred practices of data sourcing revealed some highly risky habits. Data is being collected often from the Internet and frequently originated from countries where different legislative regimes apply. While this does not necessarily make the processing unlawful, it nevertheless raises some questions about the ethics of the practice. Such considerations were discussed only by one participant. The trading of health data is a deeply troubling practice. Re-identification of data is increasingly possible, and the commercial interests around health and insurance companies, mixed with an underfunded NHS and the opening up of the post-Brexit UK market to powerful corporations, require urgent measures to protect health data. Real enforcement of the current DP legislation and more information around AI, DP and privacy would be a good and urgent starting point.

The considerations around the reduction of efficiency and the loss of accuracy of algorithms, and the low awareness of management (especially considering the increasing use of out-of-the-box systems), are concerning. The risk of repeating or forgetting the past while making predictions shows the importance of having management that is aware and well-informed and ML developers who are competent.

Surprising are the findings regarding data manipulation, reverse engineering or adversarial data which do not seem to be really on organisations' agenda. Data access to a controlled environment is another important aspect that can lead to data breaches, again, only mentioned by one participant.

The choice between a system making autonomous decisions and one that supports human decisions is particularly interesting. The latter appears to gain importance as organisations want to retain control and prefer having humans making decisions. However, noticeable was the lack of specific considerations regarding both the capacities and skills of the humans called to make such decisions.

Other risks are specifically linked to wrong assumptions made while managing innovation and lack of clarity over responsibilities, both internally and externally with partners.

4.4.2 GDPR

Organisations do not appear to have done a lot to become compliant and to fully appreciate the strategic potential of privacy-enhancing technologies and DPIAs. They can be important strategic tools for preventive DP. The case of DPIAs is emblematic of this gap. These are performed more than required in the public sector, often a sign of a tick box culture resulting from external pressure and perceived obligations. Still missing from the private sector landscape, they are often completely ignored by small entities, usually the ones more focused on responding fast to the market. DPIAs can be an important strategic instrument for organisations. Staff with different expertise in the organisation should come together and take the time to carefully examine new projects. By creating an obligation for exchanging information and fostering dialogue, DPIAs improve the organisational innovation process, becoming an important Information Management tool. This is particularly important with AI, where the involvement of different disciplines and areas is highly recommended, for example, to reduce the risks of biases. While they may be perceived by some organisations as a ‘luxury’ or ‘cost’, DPIAs are an essential tool that encourages organisations to stop and think carefully about the impact, technical and otherwise, of their innovation. This is especially important in digital transformation, where DPIAs can enable a greater range of stakeholders to be included in the process which can not only reduce risk but also develop greater organisational knowledge and understanding. Something similar emerged with regards to the DPO. The independent expert whose role is to support organisations is often not perceived as such. In the public sector, this role is disliked for its alleged power. In the private sector, it is sometimes adopted in low-risk organisations, and completely absent or ignored by small ones, many of whom are high-risk entities.

There is clearly some way to go in developing organisational awareness of the benefits of the DPO, unsurprising given that organisations are still more reactive than proactive in their approach to DP, which is rarely considered as a strategic and competitive factor. For organisations adopting AI technologies, the DPO is a potential ally and could mitigate the risk. This is especially important given the extremely rapid pace of AI innovations and their applications in very rapid market dynamics, where companies have to respond fast to the market, reducing the time available for careful considerations of lawfulness.

Another crucial point was related to the technology and the data being used for different purposes. Identifying a new use for data processed via AI, once the technology is already implemented in organisations, was one of the most concerning elements that emerged during the interviews. Identifying other purposes without careful consideration of the lawful basis is risky, deeply problematic, opposed to the GDPR approach and potentially unlawful.

4.4.3 FAT

Demonstrating compliance is a big source of concern for those more AI literate, and while awareness around black boxes is growing, other elements can impact on the capacity of an organisation to demonstrate their compliance. For instance, power knowledge and group dynamics are less considered. This can be particularly problematic in organisations that have a clear accountability structure but use innovative technologies that are not completely understood by leaders and senior managers. Similarly, while awareness around biases in algorithms is growing, the praxis of fairness can at times produce potentially opposite results for DP, as seen in the case of the increased amount of personal data collected for fear of discrimination. This could have other outcomes, for instance, on data management and security, data retention, the right to be forgotten, and the principle of minimisation.

Accountability was considered a requirement both easy, when DP was conflated with security, and extremely difficult when linked to intelligibility and explainability of the algorithms.

A different degree of explainability in decisions made by automated AI was reported as more or less acceptable according to different sectors or cases. Augmented AI was generally seen as low risk, as having a human in the loop can improve or change the decisions made by AI.

Specific problems that could arise from the interaction between humans and machines were not mentioned by any participant. However, some potential issues can result from underestimating or ignoring specific risks in Human-Machine Interaction (HMI), such as being over-reliant on decisions made by AI or making biased decisions. Furthermore, the issue of new business models based on automated systems, and their lawful bases of processing was an extremely interesting point raised only by the participant who was a privacy lawyer.

Considering the expansion of more autonomous business models, any future clarification via the courts and Data Authorities will be very helpful, also considering the rush to automation created by the COVID-19 pandemic, that is going to accelerate this adoption.

4.4.4 Key elements in the survey with experts

The key elements identified in the analysis of the data are presented in the table below:

Table 4-1 Key elements of the organisational praxes recalled by the experts

Technology	• Increasing value from data • Creating innovation is easier and faster • Loss of accuracy of algorithms (and low awareness)
People	• Human supervision and technology • Communication, exchange of information and silo mentality • General low awareness of AI and GDPR • Not many scientists in organisations • Impact of assumptions on processes (e.g., delegation and effective responsibility) • FAT and diversity of understanding • Fairness is mainly linked to data. Awareness is increasing but not into praxis (e.g., internal processes)
Processes	• GDPR compliance varies according to contexts • A holistic approach to risks • GDPR requirements/processes not used as innovation tools • New technology, new regulation, market pressure. No time for modifying processes

	• Hacking risks
Stakeholders	• Source of expertise and/or risks • Many start-ups innovating fast • Many organisations are trading health data
Decision-making	• Automated and augmented processes • Technology comes before the business case. Impact on compliance • Fat, opaque algorithms and continuous learning. Impact on informed decision making and accountability
Power	• The perceived power of DPO changes in different contexts
Innovation	• Fast innovation, time constraints and impact on sustainability and ethics • Importance of the context • GDPR is considered as good/bad for innovation according to context and personal awareness • Need for a holistic approach

4.5 Conclusion

This chapter presented the analysis of the data collected via a survey with a group of experts in AI, DP, and business technology. Firstly, it highlighted some key factors in the management of AI, such as the acquisition of the technology and the implications of DP, the importance of having knowledgeable leaders and its main risks. It then unveiled the low level of GDPR compliance, the ambivalent application of some key requirements, and the main issues around FAT, such as biases, intelligibility and explainability. The chapter then ends with a discussion of the key elements identified in the analysis of the data.

The insights from the survey proved essential for the research design. They highlighted some key areas to explore and informed strategy and understanding of the case studies. As seen in 3.6.3, the questionnaire was updated considering a less structured approach, and more attention was paid to some significant aspects, such as the role of different stakeholders on innovation, and different decision-making processes (e.g., around AI strategy and project development).

CHAPTER 5: CASE STUDY 1 – LEARNER ANALYTICS

5.1 Introduction

This chapter presents the first case study (CS1), a UK organisation creating an analytics solution with ML anticipated to perform some predictive analytics. After introducing some elements in the discussion around the use of technology in education, the chapter describes the project and presents the analysis of the data. This is done considering the three main areas, AI, GDPR, and FAT. The chapter ends with a comprehensive discussion of the key elements identified in the analysis.

5.2 Learning analytics

The use of technology in Education has grown massively in recent years. Students can now access various learning resources, learning materials, and interact with other students via social media and other online environments set up by schools, colleges, and higher education organisations to enhance their education (Kurshan, 2017). The growth of online learning, the increased availability of data and the development of technical capabilities have all contributed to the development of learning analytics. The possibility of tracking and monitoring the activities of students is growing, and learning analytics provides the capability to record, save and analyse what happens in the learning environment (Pardo & Siemens, 2014). Learning analytics systems can include ML. Past student data is used to train ML algorithms, which are then used to identify patterns in current student engagement and predict future patterns.

The interpretation of the results can lead to different consequences. By attempting to predict the future behaviour of students, the organisations can set up early interventions, offer more support, and plan a possible future of personalised programmes and education. By analysing data from the use of physical resources, organisations can change the strategic asset allocation of their resources.

Yet, concerns regarding the use of ML are growing. Organisations already collect a vast amount of personal data from students (some of which is special categories of data)

(Shearing, 2019), and learning analytics programmes come with a higher degree of invasive education, potential surveillance, and an increasing amount of data being held by organisations.

As the number of organisations implementing learning analytics is increasing, the implications for privacy, DP and ethics have received more attention (Ekowo & Palmer, 2016). For instance, the collection of digital data of young students has faced strong opposition from some parents, who are mainly concerned with security risks and risks associated with data being sold to marketers or stolen by hackers (Kharif, 2014). Several cases of U.S. vendors using data collected from systems emerged, with some of the data shared and sold to third-party vendors, such as recruiting agencies (Kurshan, 2016; S. Johnson, 2017). These practices sparked concerns over the assessments performed by some companies and job agencies, which are using employability scores based on data regarding candidates 'academic careers (Kurshan, 2017).

Another risk is associated with changes in company ownership. Whilst the initial data usage might be managed effectively, the change of company ownership can pose higher risks of data breaches. There have been a number of breaches attributed to the complexities arising when companies are bought (ICO, 2015).

A further risk is the concentration of data. Learning analytics applications necessarily involve a high concentration of personal data, and this increases the risk of attack from hackers. This risk is high, as evidenced by some attacks in 2019 and 2020: 62 U.S. universities suffered a breach (Cimpanu, 2019); an attack against Lancaster University resulted in the leak of 12,500 students' personal data (Corfield, 2019); more than 20 universities in the UK, U.S. and Canada have been victims of a cyber-attack that compromised a third party (cloud company) (Tidy, 2020); the University of California was a victim of nationwide ransomware attacks (Jablon, 2021). Such examples show the urgency of DP concerns regarding the data of students.

5.3 The Learner Analytics (LA) project

In the following, the project being implemented by the organisation of Case Study 1 (CS1) will be presented.

The Learner Analytics (LA) Project is being developed by a UK Higher Education organisation that is establishing a cloud-based end-to-end analytics solution. The activity of students in the learning environment is measured to assess engagement against the average activity of the other members of their cohort, with ML eventually anticipated to be used to perform predictive analytics. The prediction of the outcomes will inform actions to support students, providing the potential for increased retention, achievement, and completion of courses. Student personal data (some of which is special category data) is being gathered from various sources within the organisation and stored in the Microsoft Azure platform.

The current stage of the project is focusing on building the engagement system. The capacity to perform predictive analytics is expected to follow.

Student attendance is considered a key factor of engagement, and it is recorded using an App.

The organisation plans to identify students at risk, arrange appropriate interventions, and enhance retention. Students are not given the possibility to opt-out, as the project is said to be necessary to support their learning and personal wellbeing.

Further consultations and involvement with key stakeholders, such as students, have already been considered by the project board as critical priorities before implementing the predictive analytics element.

Participants were willing to provide their contribution. Everybody supported the project as that was seen as an important instrument to improve the students 'experience.

Many differences were identified in terms of understanding and perception of the capacities and the strategic role of the GDPR and ML in general, and in relation to the organisation.

5.4 Analysis

The case study was developed using semi-structured interviews, documents analysis (two business cases, DPIA, user guide, privacy policy, promotional material, website) and participant observations (more details in the Appendix B). In this section the analysis and key themes identified in the data are presented. Their specific sources are indicated by a code (e.g., CS1/1 = interview with participant 1, CS1/D = document, CS1/O = observations). The information is presented considering the three main areas (AI, GDPR and FAT). More detailed analyses of documents and observations can be found in Appendix B.

5.4.1 AI management

Different key elements were identified within AI management.

a. Strategic decisions

The project is important for business improvements, brand reputation and the economic benefits associated with increased retention. Data is viewed as being capable of providing "evidence based strategic decisions" (CS1/D), increasing organisational efficiency and compliance, optimising resources, and demonstrating a "new culture of quick delivery". The project aims at replicating existing processes and releasing staff capacity to facilitate student-facing intervention (CS1/D).

The organisation started thinking about the project three years earlier. The external pressure on the organisation and the diversity of needs of a growing number of students were reported as some of the reasons for monitoring student attendance and for using AI. The number and diversity of students have increased in the last few years, and there is an increasing need to identify and better understand their needs. Students are now more diverse than in the '80s, when the students were typically white British men, with similar and more predictable needs. Current students are more diverse, may not have financial support, or come from estranged families, or have mental health issues (CS1/8). Furthermore, data started to be very important for Equality and Diversity (E&D). The Education sector started to look at technology to better understand this complexity and support students (CS1/8).

Furthermore, the organisation is required to provide reports on impact, and it is expected to monitor student progress, making sure students are properly engaged and getting value for their fees (CS1/4). Students are also charged over the basic fee cap, and this creates an extra obligation to demonstrate student success (CS1/8). Some regulations and external bodies (i.e., professional bodies and the Home Office) already required the monitoring of the attendance of some students, such as those studying specific courses or international students.

The organisation started to focus more on the connection between attendance, wellbeing and engagement, and to monitor the attendance of all students. Attendance is viewed as an indication of engagement, and for this reason monitoring students is '*the right thing to do* ' (CS1/3). In the past, monitoring attendance was done manually, and it was labour intensive (CS1/3). Therefore, the organisation started looking for alternative systems. ML and Analytics provided some help. After seeing a product presented at a trade fair, the organisation started collaborating with an external vendor which had already created a similar solution for similar organisations. The product was purchased '*without speaking to the users or speaking to anybody'* (CS1/5), was a centralised predictive analytics tool sourcing data from numerous internal systems. This provided a different view of the students, creating a picture of the engagement via an aggregated engagement score (CS1/5, CS1/3). The product bore some similarities with the current LA, and it was piloted in two schools.

Even though the product worked, its calculations were applied universally across all schools or programs (CS1/5), without taking into consideration their differences, which limited its applications and outcomes. After the delivery of the first product, the organisation ended the collaboration with the vendor. The decision was due to several reasons (CS1/3). The product was deemed inadequate, and the (expensive) contract with the vendor was ending. The organisation believed they could create a better product managed by their in-house capabilities, while also avoiding passing on data to an external vendor (CS1/1).

The current project is being developed by an external ML developer with the support provided by Microsoft, happy to learn from a project new to them (CS1/3).

The project provides a platform to proactively engage with students, as requested by many within the organisation:

> *People started saying well actually we want some form of engagement…some form of learning analytics [...] when we can put multiple characteristics in and see if there would be an algorithm that could see if the student would attain or would drop out…(CS1/8)*

The project aim was clear, but some decisions about the technology turned out to be challenging.

b. Decisions about the AI model - Automated and Augmented AI

The creation of an automated system was the initial purpose of the project (CS1/2).

The idea was later reversed by the board, which did not want a system making autonomous decisions, and agreed that the system should not implement actions without human interventions. An augmented system was then preferred, with a human part of the process as '*a point of intervention*' (CS1/5) and as a decision maker at the end of it.

How this change occurred is particularly interesting:

> *The brief and statement of work at the outset, nine months ago, majored more heavily on the ML and AI aspect. As time has gone on and almost as I said, the pennies have dropped in and thought processes have evolved around not only*

> *Learner Analytics but the attendance monitoring plus other pieces of digital IT* (CS1/5)

As awareness about the technology increased, the focus shifted from the initial vendor, and the role assigned to ML within the project decreased. The choice to have a more human-centred system was unlikely to be reversed in the future (CS1/5). The lack of an automated decision was a reassuring factor for various participants (CS1/1, P2, CS1/7), and while some automated decisions were considered acceptable in some sectors (e.g., insurance), they were not in many others (e.g., healthcare, education) (CS1/5, CS1/7).

c. The augmented process

The system being developed was measuring the present level of engagement, using past and current data as an indicator. It had not made any prediction as the prediction capabilities were going to be added soon. The product was considered a support and flexible tool, more '*an engagement piece…for proactive engagement'* (CS1/7) than a diagnostic one. It did not aim at '*defining and steamrolling a universal set of processes, but more as a tool that SPAs*[2] *in different schools might use differently'* (CS1/5). However, SPAs are envisaged to work collectively, almost '*checking each other's interpretations…to make sure they are proportionate and consistent…*' (CS1/5).

The data of students is compared with that of their cohort. LA can identify some variations and alert the school about a drop in the engagement level. The "human in the loop" (Student Progression Administrators/SPAs) receives the alert, evaluates the data, and assesses the variation. This could be resulting from an event regarding all students, such as a general drop (e.g., teaching), or only a single one. By having a conversation with the student, SPAs can understand what caused the drop, make sure there is a real issue (CS1/8), and decide how to support the student (CS1/1). SPAs were said to have the opportunity to pause and think about the situation and the outcomes of the decision (CS1/7). Having the time to carefully assess a situation was seen as more feasible in the public sector than in the private sector (CS1/7). A human was said to have a contextualised view of the data provided by the system (CS1/8), is able to evaluate all elements, and make a decision that offers a certain level of guarantee, as '*the relevant staff can make proportionate and appropriate decision'* (CS1/5).

[2] Student Progression Administrators

Human decisions were not considered completely risk-free due to potential assumptions (CS1/2), but humans were said to monitor procedures to recognise risks and adopt qualitative processes for weighing up different factors.

Humans can also make the opposite decision to the one suggested by the ML system. If a specific situation allows it, they could use their discretional power to give the student a chance (CS1/7) which would be otherwise denied. Staff can also record data or correct mistakes if some technical issues occur in class (CS1/1).

In BAU, the project will follow the same governance practices and controls as the programme it is part of (periodical reports, dashboard updates, programme risk and issue log, joint dependency management, monthly Project Steering Board meetings). No specific requirements are envisaged for ML (CS1/D).

d. Decisions about data and the role of diversity

Participants discussed the data used by the project, and the underlying assumptions made while choosing it. The measure of student engagement is provided using circa two years of rich and various data. For example, student records and demographics, and self-declared disability and ethnicity; data that could be linked to their socio-demographic background (e.g., region of origin or home location while at the university) (CS1/2); data location proving attendance confirmed by matching to a GPS coordinate (CS1/5).

The organisation was also collecting some extra data to better support their students (e.g., non-binary identities and care leavers). This activity was informed by the work done with specialised bodies and external stakeholders (e.g., Stonewall) (CS1/2). Of notice was the lack of reference to disabilities in the user guide (CS1/D).

The selection of data for ML and the thinking behind those choices were also discussed. CS1/1 recalled the dilemma regarding new students' data. Should new students' data (received from other organisations) be considered in the prediction? This could risk becoming a self-fulfilling prophecy, with those students who had not performed very well in the past running the risk of being penalised by the system.

Other topics were related to data which is used to suggest the level of engagement in different schools, its weight in the calculation and the issues experienced with the former project.

Data is used to measure the students 'engagement against their cohort, and it is done considering the same module and the same year of study. This specificity had been underestimated in the former project (CS1/8), leading to the creation of a very basic and unsuitable system: *'the principles behind it were quite blunt in that it was parameterised, and the tool worked on quite binary algorithms, in so much as there was a right number of logins to a certain system… '* (CS1/5). The system was inadequate due to a lack of granular data at the module and school levels. Another layer of complexity in relation to student data was highlighted by CS1/9 and CS1/2, who could see how the diversity in the data mirrored the diversity of the student experience.

The importance of a deeper understanding of the link between the complexity of experiences and attendance was crucial for CS1/9, and this was also becoming more evident for the organisation. For instance, the connection between different backgrounds and experiences (e.g., white or BAME students, first or second-generation immigrants, or their distance from campus) and impact upon attendance.

CS1/9 was also aware of intersectional dynamics and could see the risks from making assumptions about students and their data. While the participant was not directly referring to their experience on the project, they were aware of these differences and knew that considering them while analysing the data was important.

CS1/2 referred to the diversity of experiences as well, but in different departments, with BAME students more prevalent in the Arts, and Asian and Middle Eastern students more numerous in the Business School and Science (CS1/2).

The capacity for making sense of those differences was also hampered by the lack of diversity amongst staff, resulting in limited representation for students, since '*BAME staff tended to be more at professor level or estates level, and actually when you look at those people who are front facing, who actually teach our students, they are mostly white. The SPAs are white…*' (CS1/9). This element was also considered important in relation to the attainment gap, as it is critical data that can provide some insights on why this is happening (CS1/9).

e. ML, past and future. Differences in understanding the technology

Different views on the importance of past data in predicting the future were held by participants. Such differences mirrored their different approaches to technology.

1. Using past data to predict the future. The possibility that the algorithms could not work well, or not as prescribed, was not considered a potential threat: *'The model is trained with actual data, and it is cross-validation. You are validating that the model is performing as expected with data that it has never seen before'* (CS1/6). Two thirds of the data is used to train the algorithm, and one third is used to check. Retraining the model was envisaged only for fundamental changes, such as a change in human behaviour or in the available data. Particularly interesting was the reason: *'The theory is that human behaviours would not really change. The theory is that the data you have it trained it on is as valid next year as it was last year…'* (CS1/6). Thus, if nothing has changed then *'it's safe to assume that historic behaviour matches current behaviour'* (CS1/6). SPAs who read the ML prediction, and make a judgement on top of it, were expected to notice the difference.

   They would apply their knowledge about the individual, or the programme, or whatever...so it's not… it doesn't really have.... ramification. It's not like it has financial implications...It doesn't directly influence the students anyway; it's not proving any bias...is just alerting pre-emptively to the people that are supposed to be looking at the students that might need a bit more attention than they have been given (CS1/6).

   No other participant mentioned the need for specific checks on ML functionality or the need to retrain the model. Some errors were expected from the new system, and additional meetings were set up within the first two months, and a review of the system after one year (CS1/8).

   The project will use unsupervised learning: '*A model will be trained on the back of that data and then presenting it with new data as we go forward'* (CS1/6). New engagement data will be fed daily into the model.

   The exact criteria suggesting lack of engagement had not been chosen. They will be based on past actions taken by the organisation in similar cases:

   We want to predict that behaviour in advance before happening […] it will be predicting whether an intervention would be raised. So pre-emptively it will bring attention to the individual student […] it is just saying that, historically, for this kind of student you would be raising intervention (CS1/6).

No new criteria were chosen at that stage.

2. The limits of past data in predicting the future. A very different approach to the relationship between ML, past, and the future prediction was expressed by another participant. CS1/2 was critical and very detailed in their analysis of prediction and expressed some doubts regarding the possibility of formulating good predictions. The ability of ML to highlight patterns in data invisible to humans was highly praised by CS1/2, who could see high value in learning from them. However, clear reservations were expressed about the possibility of obtaining good data to formulate accurate predictions.

   Firstly, patterns do not tell us if there is correlation or causation, and it is up to the human to understand the exact relation, especially if the aim is to support people or to improve or change an existing situation.

   …when we see patterns, we need to unpick the why to understand what we are going to do about it. This is where we get into that. It's telling us what has happened not what will happen, our real interest is inherent: how do we change that (CS1/2).

   Secondly, society and organisations have gone through numerous changes in recent years, and the gathered data is already old. Additionally, the socio-political climate, the employment context and the economic factors influencing students will all be different after Brexit. Would the student data informed by that be suitable to predict a very different future? CS1/2 expressed clear reservations about the complete usability of the data, noticing how learning from the past can also risk fixing it and reproducing it, and not giving the same opportunities to people who came from a different background.

   We shouldn't be making assumptions that just because somebody is the first in family or just because somebody comes from a low participation neighbourhood… we should be offering something that gives them the same opportunity for change we would offer to somebody else (CS1/2).

   Furthermore, if the prediction highlights a less successful performance, how is this going to be communicated to the student? Are they going to be treated differently from the ones predicted to succeed? Therefore, if a human is given the possibility to choose, the judgement and the action which follow the results from the algorithm can be very different:

Learning from the past is great if it allows us to make it different in the future rather than perpetuating the past… a measure of success would be the predictive algorithms being wrong, rather than just perpetuating history because we don't change our present […] the purpose of learning is that we do something different in the future. Not that we just predict that we're going to do the same thing again… (CS1/2).

Predictions provided by Amazon in the form of suggestions, or medical predictions based on biology did not create the same kind of uneasiness as automated or augmented systems making behavioural predictions that '*might reinforce a set of existing social and demographic prejudices that underlie poor performance in the past*' (CS1/2).

3. Sustainability of the projects. The organisation was considering the long-term sustainability of the project (CS1/8). Having a long-term vision meant having early discussions on some aspects of the project, for example, on available resources. The product owner remembered how these are important: *We have restructured my team in part to allow someone to come in with the knowledge to develop learner analytics. And we have SPAs already, so we are trying to utilise resources…* (CS1/8).

   By standardising a proactive process for effectively engaging people, the project will increase the availability of resources for proactive intervention, impacting on resource management. And yet, other participants noticed an increased interest in other potential benefits resulting from the project: '*there's an almost institutional benefit that can come from having a robust data set…quite interesting…and it's good because the thinking has evolved, and people have started to get an appetite for all the wonder'* (CS1/5). Thus, further uses of the technology were already being envisaged.

f. Differences in the perception of risks

In general, ML was not considered as bearing more risks than other technologies. Risks were said to be:

- Linked to the type of data, purpose of processing, staff training and security of the premises (CS1/4).

- A matter of cybersecurity (CS1/3, CS1/6)
- Not increased by the centralised system being created by the project nor by the cloud where data is stored.

Merging and gathering data into one place was not believed to carry further risks. Data was said to come from systems that have their own security checks (CS1/8) and the merging of data from different schools was seen to be '*relying on data being routinely monitored and checked'* (CS1/5). Furthermore, the use of the Cloud is considered as increasing the security of the system, because accessing a cloud environment is harder (CS1/6).

Conversely, others could see some specific risks linked to the use of ML. CS1/8 referred to an existing gap between the perceptions of the technology and its real capacities. ML was said to be less '*sophisticated'* than people think, and less transparent than people expect, increasing the risks for individuals: '*…I think we are relying on systems without appreciating the risks….You know the areas we do not know about…The bigger risk at the moment is that we are talking about individuals….about fair individual assessment'* (CS1/8).

Some interesting dichotomies in the perception of risks appeared during the exchange with two participants. CS1/4 did not express any particular concern about ML while talking about the project. However, the position shifted when the participant was asked about the technology in general. Evident concerns and clear worries were voiced about predictions applied to human behaviour:

> *You're potentially making a decision based on anticipated future behaviour which to me is…problematic because people develop, and people change. People have choices they can make in the workplace…then you could…the organisation can profile an individual and say this is what is going to happen to this individual* (CS1/4)

The risk that the past could reduce the future choices and opportunities the individuals are presented with was clearly expressed. This risk was believed to be particularly concerning in working environments where employees' behaviours are monitored. When asked if they could see a similar risk occurring with students, CS1/4 ruled it out. Students were said to be part of the organisation only for a short time. Furthermore, even if they were profiled

> *… that's possibly OK. But it would be wrong...The danger would be....a student could see that in that first year we carry out all sorts of profiling......the student is*

> *going to get a 2:2, and that tends to inform what we should or may do in two to three years 'time…to inform them how the university behaves towards a student and that is wrong* (CS1/4)

Therefore, if the consequence of profiling is not increased support, but rather the opposite, then that would be unfair, as students should be treated equally and get the same amount of support.

Another participant seemed to express a different position. CS1/6 could not see any specific risk linked to ML and the project. The technology is *'largely open-source…it's nothing specific about ML, it's just ..Yeah, it's just data security, it doesn't matter if it is ML or not, it is just data security, and methodology security*' (CS1/6).

However, they added a final consideration at the end of the interview. They could see the issue created by the rapid escalation of negative feedback. Talking about the learning capacity of ML, they referred to situations where biased data can get stronger very rapidly in environments where models make decisions. Considered easy to mitigate, this risk was viewed as resulting from insufficient data preparation.

The organisation adopted some technical measures to increase the protection, as per GDPR, such as encryption at rest and encryption in transit, and they mitigated any sort of data loss exposure. However, it was said to be 'unconsciously' increasing the risk with regard to the GDPR:

> *I think unconsciously, they are yes…I think at the moment my personal view is that a fair amount is being done on trust in terms of the DP angle from a DPIA and information governance… on the trust of the DIT [Digital IT team] and the wider project…I think by virtue of what we're delivering. It's forcing their hand to think things through and document the policies…* (CS1/5).

Particularly important is the way the GDPR is being thought about in the project, a sort of '*reciprocal learning exercise'* where *'the policy informs the project, but then the project does something which makes you think…does the policy need revising and some tightening up? '* (CS1/5). As many organisations are still trying to comprehend the law while creating projects, this was considered understandable. But the organisation was said to have increased the appetite to do more with data, as they could see the value in that. *'I think there's an* *acceptance that we can and should do more with the data…obviously in a controlled way without exposing the institution and the students to any risks…there's value in there'* (CS1/5).

g. Summary

Under AI management, some key elements were identified. For example,

- Decision-making (some key decisions were made around AI organisational strategy, the project, and the AI model).
- Technology, processes, and differences in the understanding of data and risks (mainly linked to the positionality of people).

Those elements will inform the model in Chapter 8.

5.4.2 GDPR

The following section will explore the GDPR related implications.

a. Stakeholders, processes and DPIA

Different stakeholders cooperate through a series of processes and tools.

- The Information Governance Team and the Data Protection Officer (DPO) are in charge of DP within the organisation. The DPO and the Information Governance team communicate and meet regularly (CS1/7).

  They had a different approach to DP. The DPO was more focused on discussing DP in terms of security, internal data breaches and company culture, while the participant part of the Information Governance Team was more oriented toward emerging technologies and the impact of diversity on DP (CS1/O).

- The team is small and has experienced high turnover. Some key roles were still vacant at the time of the interviews.
- A privacy group inclusive of members from different schools communicate regularly with the Information Governance Team. The group was considered particularly important by some participants, as their members were said to '*understand privacy in their schools'* (CS1/4). The group used to meet every month prior to May 2018, and it now meets every six months.

- The staff of the organisation receive periodic GDPR training (mainly online). No specific AI/ML training upskilling is envisaged (CS1/D).
- The project is using an external consultant to build the system, Microsoft Azure is providing the Cloud (CS1/5), and a GPS-based App company collecting location data.
- The product owner was in charge of creating the DPIA for the project, a standard practice within the organisation, where who leads a project is expected to take responsibility for things like the DPIA (CS1/1). Product owners are assumed, and expected, to '*know better'* than others how to perform the assessment (CS1/1).
- Once completed, the DPIA is sent to the Data Governance team which check, approve, or veto the DPIA, and consequently the project (CS1/1). The Data Governance team is external to the project. They expect to be consulted and to check the DPIAs before the start of projects to avoid any risk of a potential data breach (CS1/7). The checks are usually related to various requirements, such as legal bases, the purpose of processing, risks, data minimisation (CS1/4). The check on fairness relating to students was mentioned by only one participant (expert of DP), who was not involved in the evaluation of the DPIA of the project (CS1/7).

The requirement to perform a DPIA forces the project team to ask specific questions while completing the DPIA, "*it is crucial to this*' (CS1/4).

High risk projects are sent to the DPO for a consultation before the approval. As AI is anticipated to be high risk, the DPO is expected to be consulted (CS1/7). If the organisation cannot mitigate the risks, they will consult the ICO (as per GDPR). DPO is not expected to be involved in the post-implementation phase/BAU.

The current DPIA does not cover predictive analytics. '*That is the additional release and at that point the DPIA definitely needs to change*' (CS1/8). The reason was explained by CS1/1. '*Sometimes when you start, you don't quite know exactly how the data is going to come through. So, they're always kind of renewing and updating as we go through the different phases'* (CS1/1).

The DPIA is updated by the product owner every time there is a change (CS1/1). The product owner for Learner Analytics reported having created another five DPIAs in the same year, an activity considered to have increased their knowledge of the GDPR (CS1/8).

The Information Governance team usually become aware of the projects only if, and when, they are informed by the teams developing the projects (CS1/7). The approval of the DPIA was granted before going live.

CS1/2 expressed some reservations about this late-stage approval: '*You should do that upfront and conduct experiments based on what the DP has got to look like, rather than, right, we got this, let's make sure it's there to test at the end*' (CS1/2).

The same late-stage approval was also reported to be happening for other projects. CS1/2 could see a possible reason for this: a low-level perception of risk. The system was probably perceived to be already embedding DP requirements, with control being in place on the datasets used by the project.

While considering projects as low risk was not unusual, CS1/2 also referred to many other DPIAs being vetoed, and projects being paused by the Information Governance team. It is of interest how the DPIA was perceived differently by different subjects. For example, as a living document to be updated with the project (by the product owner), or as a single document to be finished in two to three months (by the DPO).

Of interest was also the template used for the DPIA. A standard form, inclusive of general guidance, was used for different types of projects, and the risks envisioned were standard risks and not AI risks. Additionally, the guidance appears to indicate the involvement of other stakeholders (Digital IT) that did not appear to have been involved (CS1/D).

Therefore, not only did the DPIA of the project not consider the increased level of risk due to the centralisation of the students' data, but this was also created following more relaxed rules than normally prescribed.

Some other significant aspects were identified while talking to a participant part of the Equality and Diversity Team who were involved in other workstreams, but not with this specific project:

- Performing an Equality Impact Assessment was referred to as being mandatory for some projects, and they could see no reason for the Learner Analytics project to be exempt from this obligation.

- The Equality Impact Assessment is usually performed by the Data Governance Team '*They are quite good at doing them* […] *I have worked with members of this team before…but not necessarily on GDPR…*' (CS1/9).

Nobody else mentioned the need to perform an Equality Impact Assessment for the project, nor the fact that the Data Governance team usually does it.

Another important moment regarding the exchange of knowledge within the project occurred when another participant was asked about the GDPR. Due to the high sanctions brought up by the Regulation, they did not feel entitled to give any opinion, as they were used to doing prior to the GDPR. Because of the high sanctions, the potential risk for the organisation was perceived as too high. Such a cautious approach can also have the effect of limiting the exchange of important information and impacting innovation.

b. Lawfulness

Various participants referred to a combination of lawful bases used to justify the processing of students 'personal data.

1) Contract. The contract signed by the students when enrolling with the organisation was the most referred to by various participants (CS1/1, CS1/2, CS1/3, CS1/4).

2) Consent. It was usually assumed that people give their consent for processing their data when signing a contract with an organisation (e.g., onboarding), rather than giving a separate consent declaration independently by other lawful bases (Art. 6.1 (a)).

3) Public Interest. Processing students' personal data was seen by CS1/4 as justified by two concurrent bases: for the performance of a contract, and the performance of a task carried out in the public interest (GDPR, Art. 6 (e)). The advancement of education and knowledge through teaching and research is the public task included in the Royal Charter which established the organisation. For CS1/4, this creates the obligation to ensure that the students are properly engaged. Both bases are considered sufficient to justify the processing, however, a combination of the two was seen as more appropriate. Providing personal data is therefore viewed by CS1/4 as an act necessary to receiving an education. The students sign the contract, read and sign the privacy policy, pay the fee, and receive an education.

4) Legitimate Interest was linked to consent. CS1/4 referred to it in two specific situations:

   - Alumni. The contact the organisation establishes with former students via third parties (external agencies), alumni are invited '*to give something back. Might be money, time, research, or might be advisable. So, there is a legitimate interest that but, in this case, I think we need consent in some way*' (CS1/4).

   - Student privacy policy. The document was going to be changed to include a text broad enough to cover the predictions performed by ML. Being transparent and providing the information at the beginning of the relationship is necessary and critical (CS1/4) as asking data subjects to consent to a new purpose in the course of the relationship can be complicated or impossible. The new contract for current and future students will include a broad definition that encompasses the prediction done via ML.

Time of collection is also important for lawfulness. The complexity of data and the time of its collection was an interesting point highlighted by CS1/2. Part of the data held by the organisation was collected before May 2018 and therefore under the former DP regime (DPA98). As the exact considerations had not been thought through, it was unclear if that data could be lawfully used to train ML without new consent.

A link between reporting and anonymisation of data was also mentioned by CS1/8. The reporting for analytics is generated using different reports. The personal data in the reports is anonymised and fed into predictive analytics (CS1/8). When student personal data is anonymised and used to train algorithms, it is no longer identifiable data and therefore subject to the GDPR. This process is done '*in a way that is as clear as we can be, that we will use their personal data in this way... that the learning algorithm doesn't connect anything back to them individually….*' (CS1/2).

The risk of unintentionally re-identifying some students from the data used to train the algorithms was a crucial aspect for some participants (CS1/5, CS1/8). This specific risk was also mentioned by participants when asked about the GDPR and ML. The use of students' characteristics for creating the categories used by the predictive analytics was generally considered trouble-free. Nevertheless, some uneasiness was expressed about granular data, as this could lead to potentially identifying some students bearing very distinctive characteristics. The ability to add further data sources in future was being considered (CS1/D).

c. Monitoring student attendance

The Monitoring Attendance Project is focused on monitoring the attendance and the engagement of students. The organisation is aiming at identifying students who are struggling and needing extra support. Monitoring '*needs to be reasonable and fair'* (CS1/7), and these conditions apply to all individuals, '*whoever that may be, an employee or a student'* (CS1/7).

Monitoring also means making sure students get the education they paid for, '*making sure that they're properly engaged so that they are getting value for their fee'* (CS1/4).

The app tracks the location of students when they register their attendance in class. The obligation on students to download the app was discussed by some participants.

CS1/7 was very clear in not seeing the app as an obligation on the students:

> *It would be different if students are required to download the app, but as long as they are given the choice of downloading the app…. And hopefully, the privacy statement is very clear on what is collecting and why is collecting, and students can read that and decide* (CS1/7)

Therefore, students should be presented with the option.

d. Data retention and access

Data is in general kept for five years. Some info is being kept for longer, such as degrees. The data collected and used for LA is not going to be kept and is expected to be destroyed (CS1/4). Access to data is role-based, being dependent on the access given to the specific level (CS5). For example, the SPAs can access students' *data in the school they are monitoring, while Admin will be able to have broader access and view*'(CS1/1).

This section on the GDPR showed that the organisation is aware of the general requirements around the Regulation. However, there appears to be less certainty about the question of how this related to AI technologies.

e. Summary

Under GDPR, other key elements that will inform the model were identified. For example,

- Stakeholder management.
- Processes for assessing risks.
- Relationship between technology and people.

5.4.3 FAT

a. Fairness

Differences in meaning, understanding and practices of fairness were identified in the case study interviews. Participants discussed fairness and ML both in general and specifically in relation to the LA project. Fairness in the project was discussed within the project board and project group level. The DPO and Information Governance staff were not involved in the discussion.

The DPIA was considered a crucial tool for considering fairness before the implementation and completion of projects. People are expected to ask the necessary questions, and this was seen as sufficient to define boundaries around data collection and use by the organisation during the project (CS1/4). SPAs were said to consider different elements, contrast biases, and offer people an opportunity for change (CS1/7).

Several participants referred to fairness as being dependent on context and people. As the technology can be used for ethical or unethical means, this is rooted in the ethics of people and companies.

> *It just depends on your own version of ethics I suppose…I had a conversation recently about these negative feedback loops where you've got your model […] it would be retrained on in a negative way to become biased...so that's where approaches become unethical through….poor programming really* (CS1/6)

Furthermore, another relevant element was identified during the interviews. Whose fairness are we referring to? Fairness for the individual, the entire population, or the organisation?

The case of fairness for the organisation was mentioned by CS1/3. An intervention that optimises business and saves costs was in general considered ethical, as for example, with targeted marketing, which focuses on those customers who are more likely to buy. This was also considered ethically good due to the environmental element, as '*fewer people get* *bombarded by marketing material, so I think that's a good use....an ethical use of AI'* (CS1/6).

Conversely, AI used to deceive and manipulate voters during the Brexit referendum was identified as unethical by the same participant, and a kind of project they would never agree to be part of.

As the LA project is created to support students, fairness is considered an essential part of the project. This conviction is enforced by the fact students can access their own dashboard, and they can see *'the same information staff can see…and that is the key element in terms of fairness and transparency…they are at the heart of the decisions'* (CS1/8).

A number of participants referred to bias. Not only do machines lack human flexibility, but they could also possess biases originating from input data, classification levels and demographic factors: '*With machines, we are kind of kidding ourselves that they are not biased*' (CS2). Potential difficulties in identifying where biases were located were also observed: '*How do you decide that the data is unbiased?...who is picking out the data to feed into the algorithm is going to be an individual, they may have their own biases'* (CS1/7).

The importance of innovating through questioning categories of historical data (which can reinforce classification norms) was highlighted. For example, the organisation has recently included non-binary as a gender category.

The assumption that '…*human behaviour would not really change…*' (CS1/6), was the default explanation while the possibility of its change was presented as an exception.

The power of humans to differ from predictions was mentioned by several participants, raising questions about what to do with the measurement resulting from the analytic tool, and how to do it. Would following the direction indicated by machines guarantee fair decisions, as '*a machine does not discriminate*' (CS1/4)? Or would going against past decisions and predictions produce fairer outcomes? As suggested by CS1/7, *'...your history might not say that you can do this, but I've spoken to you…and I am going to take a chance on you'*.

Unfairness can also result from an inadequate weighting of the different factors used to feed algorithms, which as previously seen, occurred in the earlier version of the project (CS1/8).

A key element of AI is the ability to make predictions based on the data. Although this is not yet implemented in this particular system, there were many references to this in the interviews. Fairness and the accuracy of predictions based on past factors were identified as critical factors, especially in the case of highly changeable socio-economic contexts (CS2). It was suggested that prediction could help the organisation with identifying future strategies rather than decisions at micro-level: '*I think that personalisation and learning will draw less upon individual predictions and more probably big strategic interventions'* (CS1/2). Fairness during the use of engagement and predictive analytics is said to be guaranteed by access by the relevant staff who are expected to '*make the right decision choices*' (CS1/5).

The tools are assumed to be fair and efficient in creating a dialogue with users and facilitating proportionate decisions made by humans. Participants commented upon various aspects of fairness in the usage of the system: '*Ethically this project seems quite sound, the AI aspect isn't really going to be used for excluding individuals but for identifying them early to pre-emptively do something'* (CS1/6). There were varied ways of characterising this. For example, one element is identifying students who are more likely to withdraw. Providing support to them would avoid a financial loss for the university, '*a less ethical way of looking at things*' (CS1/6), although conversely, the supporting action could be characterised as supporting student success.

While the application is inspiring new interventions, the need to be cautious and certain about its use was also mentioned:

> *It's very new to us…using data about people in these ways. So, I think it's important that we do believe that this tracking is OK, we can do that. There is the technology. Is it the right thing to do? Is it going to feel kind of proportionate in terms of what someone would reasonably expect that we would do with that?* (CS1/1)

Noticeable in terms of fairness, power and agency were the dynamics between students, organisation, and participants. Many participants mentioned a growing interest in data and an increased appetite for risk within the organisation.

Respondents were then presented with the hypothetical possibility of extending the ML capabilities to include a performance management tool to be used on staff. The scenario presented was unexpected, and the question created uneasiness in many respondents (CS1/O). In general, they could see this possibility as technically possible, but not feasible due to the predicted fierce opposition of the trade unions. Temporary staff/consultants were more open. And yet, when asked if they would accept having their own work or location tracked, a stronger response was generally observed, paired with some reservations and less inclination to accept. The role-reversal scenario highlighted the power imbalance between permanent staff, consultants, and students, with different degrees of protection.

Moreover, the two consultants provided other interesting insights. They both had a utilitarian approach to technology, considered to be neutral, and they appeared to be generally less concerned about biases in data. And yet, one of the two, only at the end of the interview, warned about the risk of biases getting rapidly stronger in autonomous decision-making systems, but adding that '*probably'* other ML algorithms were being designed to avoid that scenario (CS1/O). This was, again, considered only a technical problem.

b. Accountability

Accountability was mentioned by participants concerning decisions made by machines or humans, internal culture, and organisational structure.

As previously seen, accountability and responsibility were central in demoting ML by choosing an augmented model instead of an automated one:

> *…having human beings making some of those judgments we felt like it…is probably as good as a machine making them because we can bring those factors there, into accepting the fact and then being honest with ourselves about the fact that…that it will bring with it is its own bunch of prejudices* (CS1/2)

Noticeable is the assumption that the machine's measurement is faultless, and that human judgement is seen potentially as good as the machine's one, with added flexibility in making final decisions.

Another participant stressed the importance of accountability for all staff. Citing how everybody is responsible for privacy, they expressed confidence that accountability is well embedded within the organisation: '*…it's everyone responsibility. Accountability comes up through the organisation through the different schools and different departments…. the accountability goes right to the top*' (CS1/4). While this participant refers to the importance of organisational culture and staff awareness in guaranteeing accountability, any issue related to accountability in relation to AI was completely absent.

One participant viewed the project as a tool to increase accountability and transparency within processes, strictly linked to efficiency: '*…I think the accountability…there's very* *clear traceable and explainable correlations between this tool that enables these operational processes…there's efficiency gained operationally…able to remove the need for some manual interventions…*' (CS1/5). It was stated that decisions made on students will follow rules provided to staff: '*The decisions are managed between a set of processes and guidance… everything is very clear*…' (CS1/8). Accountability is the result of clear rules and boundaries for humans. Potential issues resulting from the machine's measurements are absent or assumed to be capable of being solved by human decisions.

c. Transparency

Transparency is a key element, as higher visibility and access to attendance data (by organisation and students) is considered critical (CS1/D).

Transparency was stated by many participants to be guaranteed by the dashboard, which allows students to access their own engagement data, and which is seen as crucial to fostering trust:

> *We have to be transparent and have conversations with the student ...They need to be part of the conversation. That openness is very important...being clear with them about how we're going to use it and for the student to kind of understand our side… Students need to be able to see what we are seeing…* (CS1/1)

The dashboard was seen as a tool providing transparency (CS1/8) and enabling the conversation and dialogue with individual students (CS1/3).

Transparency is linked to engagement: '*not having transparency would be detrimental to the system*' and '*this is about fairness and transparency…this is a tool for them, not just for academic staff, they can see their engagement and…that's enabling them to be successful*' (CS1/1). One participant mentioned the importance of transparency in relation to GDPR, considered the most important principle:

> *The individual knows that when he or she gives the information over to the organization or another individual what that individual is going to do with it… and I have got very clear expectations as to what you will do with that information. And that's to me that's absolutely vital…. that also means though that the organization…has explained what it is they're going to do with that information* (CS1/4)

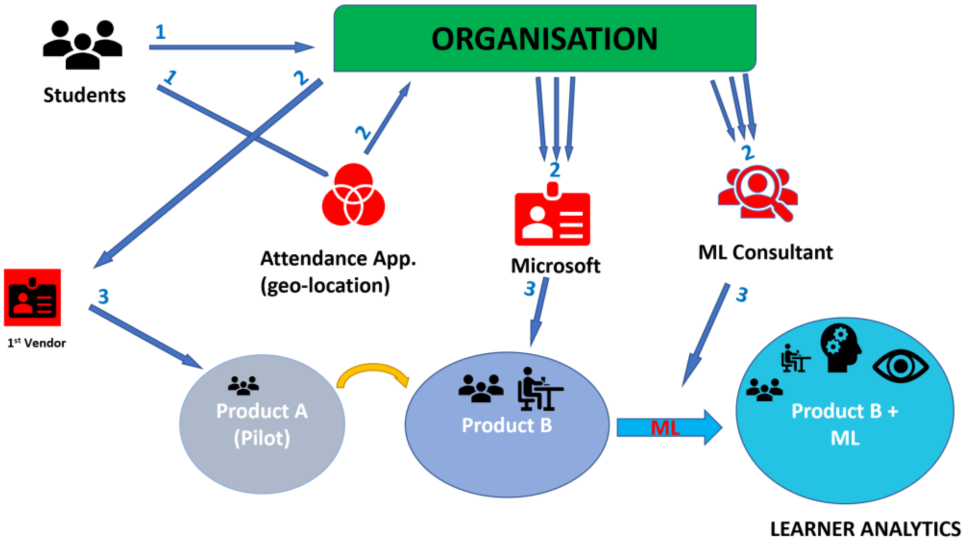
The same participant noted that transparency is strictly linked to culture through the organization, and that '*We do things properly. We understand that if we take people's data we look after it'*. Transparency in relation to predictive analytics (score and output data) was not mentioned.

This section has shown that FAT principles were high on the agenda of the project.

Therefore, the findings unveiled a complex environment where different stakeholders process students' data. As shown in Figure 5-1, data is collected by the organisation at the moment of enrolment, and this is then also processed by third parties. Additional data (location) is also gathered by another third party.

Thus, LA will then be the result of the processing done by different internal and external stakeholders.

Figure 5-1 Ecosystem of student data



(Design: Chiara Addis)

d. Summary

Various key elements were identified under the FAT principles. For example:

- How technology and FAT principles could be understood differently by people and stakeholders.
- Importance of clear rules for processes and decision-making.

These aspects will inform the model.

5.5 Discussion

In the following section, the key elements identified in the analysis of the data will be discussed.

5.5.1 AI management

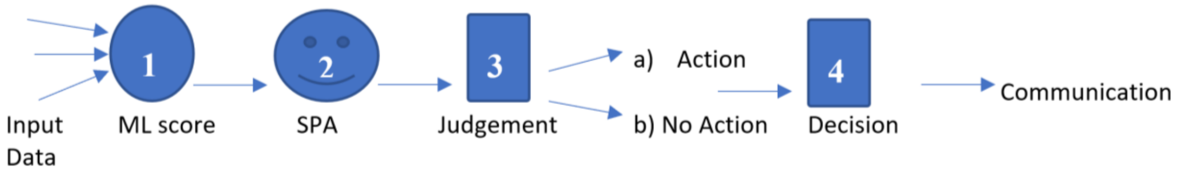
a. Human-AI Interaction: Augmented AI and the human in the loop

The way the project evolved from an automated to an augmented system is of particular interest. As the board acquired more information on ML, the desire for an autonomous system diminished. The augmented system, with its human-centred approach, was seen as providing more flexibility and more control. This change is indicative of a fairly mature organisation that questions and, if necessary, downgrades the role of ML. In the new system, the human becomes the decision maker who analyses, makes a judgement, takes the decision to meet the student (or not), verifies the need and offers some support.

However, some elements in the role of the users (SPAs) deserve to be analysed in more detail. The staff are expected to respond fast and make proportionate and appropriate decisions, as they know the context, can evaluate all elements, make sense of the ML prediction, and take a decision. There is an assumption that SPAs know what to look for in the data, how to spot an issue, and how to act, all this while being supported in their interpretations by their peers taking the same kinds of decisions.

While this straightforward process can certainly occur, the full picture could be much more complex (Figure 5-2) in terms of accuracy of the score/prediction (1), the capacity of SPAs (2), the accuracy of the judgement (3), and final decision.

Figure 5-2 The different stages of an augmented decision-making process



(Design: Chiara Addis)

For example, ML predicts the level of engagement that may impact the retention and success of the student. i.e., a low engagement (1). The SPA (2) can consider the results to be right/accurate, or wrong/not accurate. Therefore, different possibilities can occur:

1. Predictions are believed accurate. The ML score is considered correct, and the accuracy of the algorithm is not questioned. This was the position expressed by the IT project manager and the ML developer. There are two potential outcomes:

    a. Predictions are not questioned, they are believed accurate, or even more accurate than human predictions. For example, in the case of a low score, the SPA makes a judgement, acts, meets the student and sets up a support plan, helping them to succeed. Within this scenario, there is an assumption that, without that intervention, the outcome would realise the "self-fulfilling prophecy" of ML. This was presented by one of the leaders as the best use of ML when the organisation acts to change the outcome of a negative prediction. In this case, human agency alters a potential future, or to put it in other words, ML says no, SPA says yes.

    b. Predictions are believed accurate. The SPA does not act to meet the student and decides that no further support is needed. This could happen, for example, if the user does not foresee any possible improvement following a support plan, i.e., ML says no, SPA says no.

2. Predictions are not believed accurate. Mistakes are identified by a human, ML score is questioned and considered wrong. This possibility was rarely mentioned, as participants seemed to be more inclined to consider ML prediction accurate. There are some potential hidden risks in this scenario:

    a. Qualitative processes: There is awareness amongst leaders that SPAs' decisions might not be completely risk-free, due to their potential assumptions. Yet, there is confidence in SPAs adopting qualitative processes for weighing up different factors and making fair and correct judgements and decisions. It is unclear how the judgements and the decisions made by the SPAs in their interaction with the system will be checked and evaluated. For example, if an SPA does not deem some alerts about a specific student worthy of attention, and does not act to check, how will the organisation be made aware of it? Is this going to be included in a report for the line manager? This point was not discussed. Furthermore, SPAs are expected to learn from each other's interpretations (making sure they are proportionate and consistent) and the organisation wants to build a "new culture of quick delivery".

b. Experiences and knowledge: SPAs were said to apply their knowledge about the individual, and through qualitative processes, to make the decisions. Therefore, the experience and knowledge of SPAs is an important factor in making judgements and decisions. Diversity matters are relevant here. For example, the lack of representation of BAME amongst SPAs was striking. Moreover, disabilities were not a prominent issue during the interviews. This can be seen as a problem considering that disabilities can also impact attendance, a key measure of engagement.

   While the researcher does not support an essentialist approach, an increase in diversity amongst SPAs would be extremely valuable, taking into account the diversity of the student population and the significance of SPAs' knowledge in interpreting alerts, ML predictions and in deciding whether or not to meet a student. This is particularly important if SPAs suspect the scores to be biased. For instance, this could happen in an environment where historical discriminatory practices and institutional racism informed the data used to train the algorithms. The human in the loop can use their experience to spot potential bias, and their discretionary power, typical of an augmented system, to make an opposite decision to the one suggested by the algorithm. The human could give the student a chance and provide support (scenario a above) or deny it (scenario b).

   Therefore, having humans in the process who are knowledgeable about diversity issues is critical, as they could help stop the hidden reproduction of historical biases, a massive risk with ML.

3. Training: The planned staff training will be internal training. No specific AI training was foreseen in the business case or mentioned by participants. Thus, anything specific on the interaction with AI/ML, and on how to identify potential technical issues or loss of algorithmic accuracy had been planned.

4. Feedback data: The behaviour of SPAs will have an impact on future learnings of ML as ML keeps learning from the environments where it is deployed. The interactions (or the lack of) between SPAs and students will create further data. This will become a new "past", to be used as feedback data for the future learning of ML algorithms.

Similar risks could be evaluated when performing DPIAs and equality assessments, but the assessors need a systemic/holistic understanding of the organisation.

b. ML Predictions. The role of past data in predicting future behaviour

The role of the past in predicting future behaviour is a crucial element in the creation and use of ML by the project, which is said to enable evidence-based decisions. However, very different approaches surfaced during the interviews:

1. <u>Predicting the future from past data?</u>

   The position of the participant creating the system was clear. Retraining was not considered necessary as human behaviour would not change. As past behaviour is expected to repeat, retraining is expected only in the case of a critical and fundamental change in behaviour.

   This is a significant point, which can lead to different situations:

   a. Who decides the criteria for triggering the need to retrain algorithms? This is not clear. Human interventions are also shaped by the dynamics of power. It would appear that SPAs are expected to identify the change in students' behaviour, generally read as a potential need for more support. Are SPAs also expected to recognise a critical change? For example, what differentiates a temporary drop from a fundamental change? Would that be a numerical one? For example, the higher (and more visible) the number of students changing behaviour, the bigger the change needing to be transposed into the model. Will small and less visible changes (with a potential for big impact) be considered? What is the process after that? It is unclear. Similarly, loss of accuracy or the possibility that the algorithms would not work as prescribed were not considered.

   b. System sustainability: The type of ML being implemented (unsupervised learning) looks for patterns in data, does not learn by predefined categories and works with minimum human supervision. This was not necessarily viewed as needing extra checks. The options for testing and controlling the stability of the algorithm over time were not mentioned. It was unclear if the person in charge of building the system was also going to check/monitor the results for a certain amount of time after the implementation. While a series of meetings was planned for the first period and a review of the system after one-year, specific checks on the ML functionality were not mentioned, nor included in the documents.

The long-term sustainability of the system was being considered, with discussions on some aspects of resource management and a plan to move staff to facilitate intervention. While the sustainability of the system might have been left for later consideration, it was significant that ongoing performance and technical sustainability of the system (with checks, tests, and periodic controls on algorithms) did not feature in the data collected.

2. Fixing the future by replicating the past?

More critical stances on the general capacity of ML to predict future behaviours were expressed by participants who did not have technical roles, for example leaders. While ML was considered good for identifying hidden patterns and useful for strategic planning, some reservations were expressed regarding different aspects of ML. For example:

a. Input data: Extremely fast socio-political changes risk making training data quickly obsolete.

b. Predictions: Correlations can be mistaken for causations. ML can reproduce past mistakes, creating the risk of obstructing the opportunity of individuals who share some characteristics with people excluded in the past.

c. Post prediction: It is an open question as to how to communicate a negative decision to the subjects affected by it.

Some of these observations pose some interesting questions for the management of the organisation. For example, is the result of the calculation (the ML prediction) going to be communicated to students who are going to exercise their data subject rights, e.g., Subject Access Request? How is the organisation going to respond? The decisions will not be made automatically by the system, and the provisions included in Art 22 GDPR will not apply (as seen in 2.3.6). However, that score is personal data, and it should be communicated in response to a Subject Access Request. This point remained unclear as the DP details for the post-implementation phase were not mentioned.

Certainly, the observations made by one of the leaders show a high level of awareness and knowledge of ML and its potential implications for the context where it is deployed, a deep level of understanding that appears to be growing with the evolution of the project and the acquisition of new information on the technology. This is significant and demonstrates the importance of having leaders who understand different data, its impact on data subjects and contexts, who are aware of the structural dynamics of social exclusion, have a holistic vision of the organisation and its responsibilities, and are not afraid of changing the role of the technology when believed necessary.

c. Quantity and quality of the information in student data. The case for diversity

The organisation holds large and diverse amounts of student data. The datasets used will include the last two years of personal data, including special categories of personal data (such as disability and ethnicity, and socio-economic origin). The organisation is already asking students for extra data (e.g., care leavers and non-binary gender) in order to better understand their specific needs, and it is already thinking of gathering further data in the future. The concentration of data is supposed to increase knowledge, allowing for the prompt identification of students 'needs and supporting the best action in specific cases. Therefore, the more data is available, the higher are the chances that ML can identify hidden patterns between data, providing predictions.

And yet, something more was identified in the analysis. Is having more data enough to reveal useful information on students? Does an increased quantity of information equate to an increased quality of information? Is the system able to provide that complexity in the results? Is the user of the system also capable of seeing and reading qualitative aspects and nuances, and consequently giving significance to the complexity of the information included in the data? The topic of diversity and its connection to data is important to explore some of those nuances.

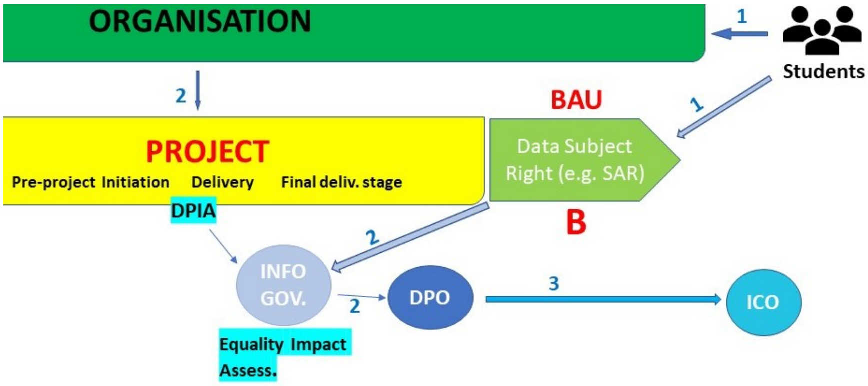
1. Diversity in schools. The previous project (pilot) appeared to have insufficient data on diversity characteristics from the various schools. The new one (LA) uses more complex data and takes differences between the schools into account. Does that include, for example, BAME students? And if it does, does it consider that more black students attend Arts, while more Asian students study Business? Does it consider immigration status? All those elements can affect the attendance of the students in different ways (CS1/9). If such granularity is not provided by the data, SPAs should

be able to address those peculiarities with their understanding while analysing the data. While this is obviously advisable, it could also create unfair interpretations and practices if no clear guidance is given.

2. <u>Lack of diversity in representation</u>. SPAs were reported to be all white. Again, the lack of representation could cause some issues on attendance, on the interpretation of ML scores, and the explanatory encounter between students and SPAs.
3. <u>High complexity in student experiences vs low complexity in student data.</u>

   The conversation with the E&D manager highlighted a crucial element. The E&D manager was well aware that some systems of exclusion associated with some backgrounds and experiences could impact on attendance. While the organisation was becoming more aware of such a connection, this did not appear to be translated into project practices. For instance, Equality Impact Assessments need to be performed for many projects, and this is usually done by the Information Governance team. Such an assessment is also required for the LA project, but that was not performed, which appears to confirm a silo mentality and the need for a different approach to internal processes. Thus, the role that Information Governance could play in this and similar projects would be central (Figure 5-3).

Figure 5-3 DP roles in the innovation process



(Design: Chiara Addis)

The understanding of both aspects — DP and Equality — makes their role and responsibility even more important within projects where AI can impact very rapidly on multiple aspects of the student experience.

The project board and team are already very careful when discussing the priorities of students and questioning the use of data. An example was the discussion about which data should be used as data feeds, and whether it was appropriate and fair, and also the discussion about using the grades awarded by other institutions for the predictions. Similar discussions are an example of the attention the project is paying to the well-being of students and ethics in their choices. This is, again, indicative of a mature organisation that is considering different elements while creating innovation.

However, there are some insights and good practices in other areas of the organisation (e.g., E&D), which do not appear to have been considered enough in the project. An increased involvement would be highly beneficial and would increase expertise. That would be particularly advisable as the organisation plans to expand the type and the quantity of data processed in the future.

d. Management of risk - risks, positionality, and power

The perception of risks varies considerably when discussing the technology in general, in relation to the project and according to the role of the participant. The risks linked to personal data were perceived as typical security risks (e.g. physical, and electronic access to data). Cybersecurity was a concern, but the use of the cloud or the centralisation of data were not. Similarly, the security checks performed on the different sources of data feeds were considered sufficient, and no extra checks were planned on the system after the aggregation of data. According to the process identified in the DPIA, risks are supposed to be identified and mitigated by the business owner, after discussing the project with Digital IT and Information Governance. An engagement with Information Governance (from the start of the project) in order to mitigate the security risks is also suggested in the business case. These close engagements apparently have not occurred. The recent cases of other universities being hacked, with the loss of students 'data, were not discussed by any participant.

Some surprising mixed positions were also expressed. Confidence in ML was displayed when discussing the project, but uneasiness appeared while talking about some aspects of ML in general. Similarly, confidence was shown when discussing ML applied to some people, but uneasiness was expressed when applied to others. For instance:

1. The ML developer was concerned with the ML's capacity to rapidly escalate negative feedbacks and reproduce biases in autonomous systems, but the risk of ML in the

project was viewed as a matter of cybersecurity. The rapid escalation of bias in augmented systems did not seem to be a concern.

3. A permanent member of the staff, a senior manager with a technical background, did not consider extra checks on merged data to be necessary, yet they believed some unknown aspects of ML were potentially able to impact transparency and fairness in the assessments.
4. The DPO considered acceptable the monitoring, profiling, and prediction of students ’ grades, but saw it as highly problematic if a similar system were used to monitor, profile, and predict employees ’performance. This is strictly linked to positionality and dynamics of power, both vertical (institution vs students) and horizontal (permanent vs temporary staff). While profiling was considered in some cases ethically problematic, monitoring of students was considered acceptable. While students are not given the possibility to opt-out (as the project is said to be for their wellbeing), the hypothetical possibility of creating a similar system for staff produced strong and different reactions. And yet, this also revealed a further risk linked to the management of AI technologies. If the project aims to support a group of people, for example, students or staff, and its success is directly dependent on the collaboration of the group who provide the data and trust the system, then a wider and more effective interaction with the group would be advisable for guaranteeing the success of the project and the long-term sustainability of the innovation.

Power dynamics amongst stakeholders could impact on innovative projects and the perception of fairness, especially when external stakeholders are involved. Furthermore, according to the specific company culture in the organisation, similar situations can have multiple perspectives and interpretations. For example, staff resistance can be read either negatively, as resistance to change in Digital Transformation and Change Management, or rather positively, as internal resistance, increased agency, and empowerment of employees. The latter could foster more active engagement and more participatory and inclusive practices of innovation management.

The company culture of the organisation appeared to be quite multifaceted. Constrained by a regulated environment, the organisation, its leaders, and managers displayed a certain amount of care in their innovation praxis and attention to students. And yet, the impression was of an organisation at times more focused on completing some processes than on reflecting on their significance, their value, and potential improvements. Student engagement and its importance were central to the project and were discussed continually during the interviews and in the documents. However, the engagements within the organisation, between various subjects who could provide some important contributions, appeared to be often overlooked.

All the above shows the existence of:

- Some doubts about AI in general, which dissipate in relation to the LA project.
- An over-reliance on the technical and organisational measures adopted by the project and the organisation.
- An over-reliance on the human in the loop, their control on output data and their final decisions.
- A different perception of risk according to positionality and power.

Therefore, the current perception and management of risk could benefit from a review of the wider management of ML within the company.

Furthermore, the GDPR requires organisations to implement the technical and organisational measures necessary to guarantee secure processing and to demonstrate them (Art. 24.1), as per the accountability principle. This calls for a change in the management of DP, with active involvement of the DP roles in the management of risk and all innovation processes.

Checking data (which, how, why) only at a specific moment — usually for approving the DPIA — is hardly enough. Additionally, continuous checking and monitoring of the data used would also be desirable, considering the possibility of rapid changes in the system.

The key elements — i.e., which data, how it is processed, and why — should be extended to the full product cycle. This could include the following steps:

- Increasing knowledge and awareness of how AI and DP can play out in that specific context. When this knowledge is not already in-house, it should be acquired externally. This also implies considering AI as a technology that requires more than just technical competencies. Employing a consultant to create a new AI system is not enough, as such technical capabilities need to be managed.
- Establishing mechanisms for internal knowledge transfer would be worthwhile.

- Involving, as much as possible, the end-users who will be directly affected by the predictions would also be sensible.

5.5.2 GDPR

a. Acquiring the technology and relationship with third parties

The reasons and the way the technology is acquired are important. The organisation needed a better system to respond to internal and external needs, and ML was chosen to reduce complexity and respond fast and more efficiently. The first product was created by a third party, chosen after seeing it showcased and without internal consultation. The end of the contract created the opportunity to reconsider the collaboration before the renewal, and to evaluate the capability of internal resources.

The above process includes some interesting elements:

1. The choice of the vendor was made without involving internal stakeholders. Considering the existing internal expertise, their initial support at the beginning of the project might have helped to identify the issues.

2. Creating the system in house was said to increase DP as data was not processed by an external vendor. And yet, while the new project offers more guarantees, it does not exclude the sharing with external entities, such as the GPS-based Mobile App collecting location data of students, the consultancy helping with the creation of the ML capability, and Microsoft. The details of those contracts were not shared, and it was not possible to know if shared data is anonymised and kept for research, as this appears to be common amongst some vendors who are using customers' data to train their algorithms.

   The student privacy policy can provide some insights. Allowing a certain degree of flexibility can create some issues in relation to ML and DP:

   - "Third parties *may* use the data for the exact purposes specified in the contract". Processing for another purpose seems to be implicit. This would be a breach of the GDPR, which carefully disciplines the relationship between the parts.

   - Data is either deleted or anonymised at the end of the contract. Considering the risk of re-identification, de-anonymisation, and the trade of student data, a complete deletion would be preferable. This would be sensible, especially in the case of data related to students' political beliefs. The reason for including and processing such data remains unclear.

- Data is shared with parties who have an interest in tracking student progress and attendance. While this seems to refer to current or potential employers to confirm details of progress and attendance, the specific purpose should be better clarified, considering the growing attempts to access student data (as seen in 5.2).

b. DP process

The DP "core" actors include Information Governance, a small team that has gone through several changes; a DPO, who is an internal legal expert who works on DP once a week; a less powerful Privacy Group (the internal group of "experts") who meet twice a year. No other roles seem to be involved with DP on a regular basis. Staff received online GDPR training, and no specific training on AI was being considered. DP staff become aware of a project processing personal data only when informed by its product owner, who is in "charge" of DP within the project. No specific training was mentioned and no explicit experience in DP seemed to be required for the role. There is a clear separation between the DP governance roles and the project. This demarcation guarantees effective control of the data processing activities planned by the project. However, this separation also means that the information between the two areas is exchanged with more difficulty. Specific expertise appears to be confined to separated areas. Information is provided in response to procedural needs, and verifications are conducted in order to progress to the following project stage. The expertise of DP staff is sought for assessing compliance of decisions already made in the project.

Moreover, their intervention in the project was requested later than indicated in the DPIA form and approved before going live. This appears to be a common occurrence. One of the leaders was clearly aware of the limits of this late intervention, explained as an underestimation of risks as the controls on data were already done on the datasets.

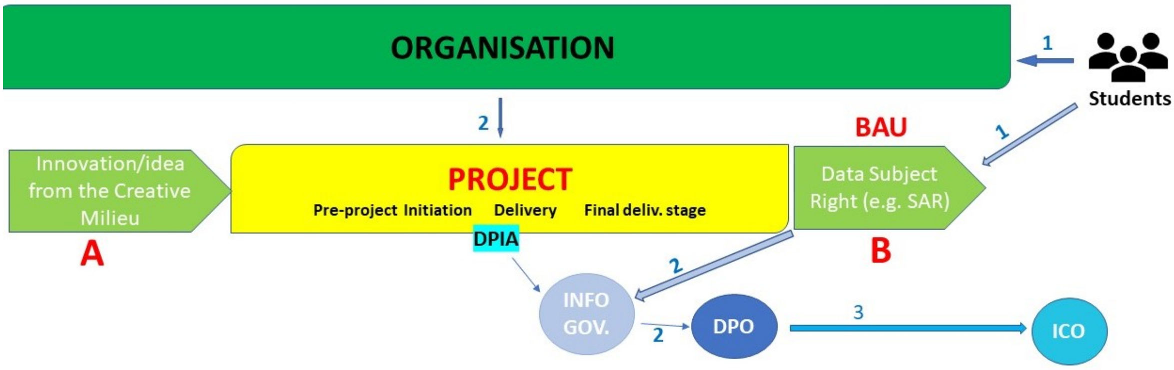
Data merging, the use of algorithms, and the centralisation of information are not considered to increase the risks, as also demonstrated by the decision to have the project subjected to the same governance practices and controls as the programme. This underestimation can be highly risky and costly for an organisation developing AI projects. Information Governance, following the DPO's indication, had already vetoed other projects whose DP measures were deemed insufficient to protect data subjects 'rights. The use of ML is generally considered highly risky, and for this reason, organisations must perform a DPIA. The approved DPIA is the assessment of the first release of the project. The assessment of ML will be included in the updated version to be updated by the product owner. There is no plan to change the DPIA template considering AI.

Therefore, Information Governance and DPO do not appear to have fully assessed the risks of ML. This can be problematic. If the risks from ML are considered too high, and the product owner's planned mitigations are insufficient, the project could be vetoed. That would be a costly issue considering that an ML developer is already working on the ML capability of the project.

c. DP staff as actors of innovation?

The activity of the DP roles in relation to the projects appears to be mainly limited to the control performed on the DPIA (when requested by the project). However, the scenario created by the interrelation between DP roles, project and organisation appears more complex than that, as shown in Figure 5-4.

Figure 5-4 The project within the organisational creative milieu



(Design: Chiara Addis)

DP roles do not appear to exercise any influence in the pre-project phase, and similarly, they do seem to be distant in the post-project phase. This can be challenging for organisations dealing with projects involving AI, as the DP is relevant throughout the full internal innovation cycle, from the ideation to the management of the product done by BAU.

For example:

1. Creative Milieu A. DP as a box-ticking exercise or an opportunity for innovation?

DP staff are not part of the activities happening prior to the creation of the project. This comprises the organisational creative milieu where innovative ideas are born and then translated into a project. While the LA project is considered an essential part of a more comprehensive digital strategy undertaken by the organisation, the DP is not considered a proactive part of such a strategy. Involving DP staff nearly at the end, before going live, is symptomatic of the significance given to DPIAs within the organisation, considered more as obstacles to remove and a box-ticking exercise of decisions defined elsewhere, than an opportunity to incorporate in the innovation process. The work done in creating the DPIA was appreciated as it offered the opportunity for pausing and thinking about the technology. Doing the assessment with the active intervention (and co-creation) of Information Governance, supported by the DPO, would have helped the product owner, and the project team, to further deepen the analysis and expand the outlook. Information Governance and DPO usually intervene "a-posteriori", after the idea is already translated into a project. They play no part in the "a-priori" activities (A in Figure 5-4). They do not appear to participate and seem not to have any influence upon the strategic decisions made by the organisation, and they have no part in the internal process of innovation. This is not unusual for similar organisations.

However, having some space where different roles/areas can exchange information, and involving the DP staff in an early stage of the innovation process (prior to the creation of the project) might have helped to facilitate the knowledge exchange, and the identification of risks, mitigations, and opportunities. That could help to integrate DP into the full AI lifecycle and to foster a more responsible innovation process.

2. Data Subject Rights B.

After delivering the product, the system will transition into business as usual (BAU), and students will be able to exercise their data subject rights via Information Governance (B in Figure 5-4). Limited involvement of the DP roles with the project can translate into a partial knowledge of the technology used. This can impact on the quality of the response to subject rights and affect the GDPR compliance.

5.5.3 FAT

a. Fairness

Different understandings and perceptions of fairness were present within the project, and the context, roles and internal negotiations were important in defining values and boundaries.

Discussion on fairness was limited to specific spaces, did not always happen at the beginning of the project and was dependent on specific stakeholders.

Fairness was mainly discussed in relation to biases and input data. Unfairness and biases in the surrounding processes and output data was not usually taken into consideration. Bias in both data and ethical use of the ML were recalled while discussing fairness, and some differences between technical and non-technical roles appeared. For instance, the prediction of future behaviour from past data is seen:

- Not to be an issue, as human behaviour does not usually change (ML expert, technical role).
- To be potentially a major issue as it would reproduce past unfairness (leader, not technical).

Similarly, a participant with ML knowledge could see the crucial importance of better programming to stop biases. Yet, they also considered targeting marketing as ethical because it is more environmentally friendly.

The leadership appeared to have a multifaceted understanding of the concept. The first leader, who had a deep understanding of bias, raised an important question. Who decides the criteria for what an unbiased piece of data looks like? That is a matter of power, and the leader was aware of the implication for AI management. The second leader identified business optimisation and cost-saving as examples of fairness and raised an interesting point. Fairness for whom? Only students or organisations as well?

Neither leader had a technical background. Of note was also the work done through questioning and expanding existing categories, for example, by collecting a further set of data for a new group identified as needing more support (non-binary people). This was an interesting example of internal innovation done using personal data. By diversifying and increasing the amount of data, the organisation aims at increasing inclusion and fairness for students who are more at risk of being marginalised. More data means more inclusion.

Furthermore, considering that this data is a special category of data, compliance and ethical use by the organisation is crucial. This is yet another example where the involvement of DP roles would enhance ethical innovation.

However, despite this, DP roles were not part of the discussion on ethics. Moreover, the DPO considered the DPIA as a document sufficient to define the use and limits of processing. This can be problematic. The same DPIA template is used for all projects and does not take AI into account. This can translate into a check done without analysing the specificities of ML, which could risk leaving potential unfair consequences undetected.

Another critical point is the assumedly "inherently" fair nature of the project. As the project is aiming at "doing good", the fairness of the project is taken for granted, with human judgement and actions in general assumed to be unbiased, and capable of resolving systemic issues. This is risky. Pursuing an original positive change cannot exclude a later unethical use of the system, for example, the exclusion of under-performing students to reduce the risk of financial loss for the institution. In such circumstances, protective measures (e.g., privacy/ethics by design) could lower the risks and indicate DP maturity.

Similarly, an ethical aim does not eliminate the risk of a project causing unfair/unjust consequences during the project cycle or in a specific environment (e.g., the pilot project).

Finally, the researcher aimed at exploring how ethical dilemmas and trade-offs were identified and discussed at different levels (e.g., board meetings). While some data provided some information, this did not provide the depth sought by the researcher around more philosophical aspects in decision-making processes.

b. Accountability and transparency

Two points are of primary importance: demotion of ML and differences in understanding of participants from different areas.

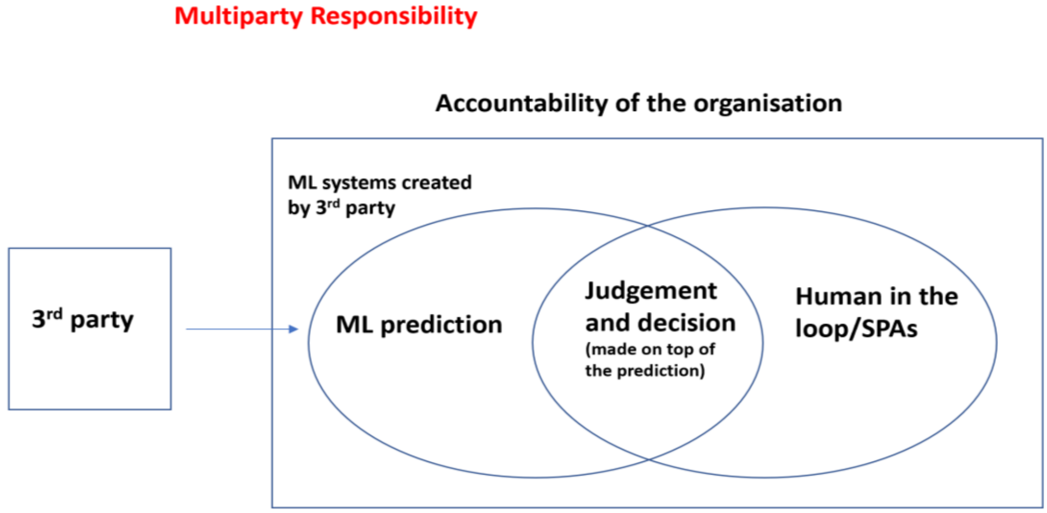
1. Demotion of ML was done to maintain the control of decision-making. The organisation did not want to be responsible/accountable for decisions made by an autonomous ML, which was not trusted and difficult to control.

2. Accountability was perceived differently by different participants, whose understanding was usually informed by their specific area of competence and their positionality. For example, accountability was seen to be directly dependent on:

- People and to be everybody’s responsibility (DPO). Thus, as layers of internal accountability exist, company culture and organisational structure are fundamental.
- The full system (technical consultant). The project reduces manual work and increases efficiency and visibility.
- The processes (internal senior manager-technical), and rules and processes for staff are clear.

Both parts (1 and 2) help to highlight something new in the debate around the GDPR and the principle of accountability: the complexity of the concept when third parties are creating AI systems. According to the GDPR, organisations have to demonstrate that they are compliant, and the responsibility is with the organisation processing personal data. In shifting from an automated system to an augmented one, the accountability/responsibility of the algorithm formulating predictions seems to disappear. Where does the responsibility lie when third parties are creating an ML system for the controller, and then leave? This is not a simple situation where a third party (processor) is processing data on behalf of the controller, nor a situation where two entities are deciding the means of processing (joint controllers). Even though algorithms are not making decisions, humans are making judgements on top of ML predictions. Does the responsibility completely move onto humans who make the decisions, and onto the organisation? From a GDPR point of view, the organisation is accountable, either with an automated or augmented system. However, could we add a different reading, emphasising multiparty liability? From an AI management point of view, is there any residual accountability attributed to the machine in the case of augmented decisions? And to the third party which creates the ML system? If the decisions are made by humans, on the basis of ML predictions, does that imply a total shift of responsibility? This is a critical point for the management of third parties and AI management, that needs to be considered by the management and clarified in the contract with third parties. This could be an example of how FAT can drive practices. While the organisation is the legal entity accountable for the GDPR, the situation is more complex from an AI management point of view (Figure 5-5). When algorithm and third parties are involved, it would be rather advisable to clarify the exact amount of accountability linked to the algorithm functionality, and the amount resulting from the human decision.

Figure 5-5 Accountability and multiparty liability


Multiparty Responsibility
Accountability of the organisation
ML systems created by 3rd party
3rd party
ML prediction
Judgement and decision (made on top of the prediction)
Human in the loop/SPAs

(Design: Chiara Addis)

This is especially important in the case of systems changing over time, and a third party working with the organisation only for a short time (as was the case with the pilot project). Similar situations seem to make the case for a co-accountability between parties managing AI systems, which would differ from the joint-controller/controller-processor GDPR regime.

Finally, the views expressed in point 2 all refer to important components of accountability and the ways it could be increased within the organisation. Yet again, the multiplicity of points highlights the need to improve the collaboration and the knowledge exchange between different areas of the organisation, and to coordinate a cohesive action aimed at improving its general accountability.

To sum up the section, transparency was strictly linked to fairness. Increasing transparency is a prerequisite to having a fair interaction with students and gaining their trust. The dashboard is the key to starting the conversation with the students who are said to see the same information that users can see. However, that does not seem to include the LA part of the project, as the predictions and the criteria SPAs will follow were not mentioned. Even in the case of mere profiling (not involving a decision made via automated means), the organisation is required to provide "*meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject*" (GDPR, Art. 13.2 (f)). It was not clear when and how this will be provided to students.

Furthermore, students were said to be reasonably expecting this processing as they are informed about it when enrolling. However, this level of granularity of data does not seem to be communicated to students when registering, nor in the contract or in the privacy statement, which is a problem.

5.5.4 CS1 Key elements

The key elements identified in the praxis of CS1 are presented in the table below:

Table 5-1 Key elements of the CS1 praxis

Technology	• The amount of data is growing, so is potential and challenges • ML – growing potential and different understanding of the technology • Relationship between technology and people, different understanding, and expectations
People	• Roles, and multifaceted understanding of AI, GDPR and FAT • Diversity of experience and background
Processes	• Full innovation process, also considering the external and internal contexts • Risk assessment processes (e.g., DPIA, Equality) • Importance of clear rules • Augmented AI process
Stakeholders	• Stakeholder identification and management
Decision-making	• Decisions around AI strategy, the project, and the AI models (automated and augmented)
Power	• Positionality • Impact on the implementation and use of the technology

Innovation	- Holistic approach

5.6 Conclusion

This chapter presented the first case study. The organisation was implementing an analytic solution aimed at improving the engagement and retention of students.

Firstly, it introduced the main elements of the current discussion around the use of technology in education, and the main implications of the growing availability of data (i.e., cyberattacks). It then illustrated the main characteristics of the project and the analysis of the data. The chapter then ended with a comprehensive discussion of the key elements identified in the analysis, highlighting, for example, the role of people in Augmented AI models, the importance of diversity, and how agency and power of stakeholders can affect the implementation and use of the technology. The discussion further stressed the crucial importance of holistic approaches to innovation and the multifaceted understanding and practices of FAT principles.

CHAPTER 6: CASE STUDY 2 – DIGITAL IDENTITIES

6.1 Introduction

This chapter presents the second case study (CS2), a small start-up (DIP) developing a project around digital identity. After introducing some elements in the discussion around identities (e.g., different traditions and main legislation), the chapter presents the project and the analysis of the data. This is done considering the three main areas, AI, GDPR, and FAT. The chapter ends with a comprehensive discussion of the key elements identified in the analysis.

6.2 Identities

The debate on digital identities is lively and complex. This section will address some of the key elements of this debate, focusing on the differences between civic and digital identities, and their role in different traditions and different legal systems.

6.2.1 Civic and digital identities

The conventional meaning of legal identity is usually linked to civic identity, recognised and verified by the State via ID documents (such as passports or driving licences). Digital identities convey a more complex nature, being often the product of offline and online information about an individual. They can result from the combination of different data linked to civic identity held in different systems. They can often encompass further information for supplementary aims, for example assessing risks in the case of credit rating, or entitlement to services, or include information collected from online activity (digital footprint).

6.2.2 Different traditions

The differences between the UK and other countries in handling digital identity systems are the result of different traditions in perceiving, managing, and regulating identities. Legal systems based on civil law, which are the result of Roman law and Napoleonic and German traditions, are typical of countries in continental Europe and are based on a codified primary source of law (Helmholz, 1990). Typical of such systems are written constitutions and identity cards. Common Law, deriving from an Anglo-Saxon tradition, is typical of many commonwealth countries and is based on legal precedents. Identity cards are not common. Continental European countries have national identity schemes, while the UK lacks similar systems, having instead public registers commonly used to prove someone's identity.

The attempt to create a UK national identity scheme, the Identity Cards Act 2006, inclusive of a national identity smartcard (Beynon-Davies, 2011) proved unsuccessful. The scheme was strongly opposed as the many categories of data (especially biometrics) included in the planned National Identity Register were considered to be a means to monitor citizens (Doshi-Velez et al., 2017).

6.2.3 Meaning of identity

The concept of identity has been analysed according to different approaches: an essentialist approach to identity (often in continental Europe), versus a more constructivist, fluid, or performative approach to identity. Characteristic of an essentialist paradigm is the idea that "…our identities (or major facets of them) emerge from the inside, and are fixed over time, innate" (Cover, 2015, p. XI). This deterministic view of identities includes the genetic, which considers personality attributes as genetically shaped. This view is at the core of some of the examples discussed by Kosinski et al. (2013) who claim that ML could be used to identify gay men and lesbians from their faces[3].

On the other hand, there is the idea that identities are socially constructed (Burr, 2015), or are the result of performative acts and reiterations (Butler, 2011). These approaches translate into a less deterministic view of identities, which creates more interpretative spaces for questioning. However, who has the power to create and define categories or to decide who is included within them? Within the realm of digital identities, and especially within the context of the exploitation of personal data by capitalism's surveillance systems (Zuboff, 2019), answering those questions is not easy. Online identities have become something we cannot ignore. As noted by Cover: "Identity is always online. We are, in some ways, always performing ourselves online…. because we leave traces all over the Internet, …[on]…sites that are actively contributing to elements of our identity… our identities are 'always on'" (Cover, 2015, p. X). As rightly pointed out by Feher, in relation to digital footprints, the question is not anymore "Who am I" but "Who am I online?" (Feher, 2019). In the context of digital identities, how do people represent themselves and how do others see them? Is the

[3] N.B. Kosinski et al. focused on gay and lesbian faces only and not on LGBTQIA* faces.

online identity a “true” representation? Is the online self the result of an active and aware choice by individuals? Are people able to shape their online narratives? If we assume they can, we also need to assume that they have power, agency, and control in creating their online identities and in controlling the potential effects these representations have on their reputation. Yet can people really control all of that? It could be rightly argued that the agency and control individuals have is limited by the growing power of Data Brokers and their use of personal data. For example, the role of Credit Reference Agencies (CRAs) has grown massively in the last few years. Created to assess the creditworthiness of customers and prevent criminal activities, they have become powerful Information Brokers with interests in different areas, such as marketing and digital identity. By gathering huge amounts of personal and non-personal data, and aggregating data from different sources, they have created databases of millions of people.

Privacy and DP academics and activists have highlighted the perils coming from the exploitation of people’s data (European Data Protection Board, 2019; Ryan, 2018),

and have often faced a certain laissez-faire attitude from some of the DP Authorities.

The GDPR compliance of Data Brokers ’practices has been questioned, for example, for breaking “the principles of transparency, fairness, lawfulness, purpose limitation, data minimisation, and accuracy” (Privacy International, 2018).

The increased need to know more about customers has fuelled the market for personal data. Many companies are reported to be trading data by buying and selling information to “clients who want to better understand users” (Murgia & Harlow, 2019). The system of ML models has also been questioned with regards to both their technical efficiency and unfair practices (ibidem), and the use of AI in B2B relationships (Gligor et al., 2021). A further complication relates to the privacy risks of inferences, assumptions or predictions drawn from the data, often without the awareness of individuals, who have “little control and oversight over how their personal data is used to draw inferences about them” (Wachter & Mittelstadt, 2019, p. 6).

The commercialisation of personal data, which is the core of this ecosystem, and the doubts about the real accountability of Data Brokers, complicate the premise of individuals’ agency and control. Do people really know how their identities are constructed and represented online? Emerging technologies are continuously collecting data from the environments where they are deployed and continuously shaping and creating new content. This is particularly the case with ML systems that are increasingly becoming more autonomous and with business models based on the growing use of third parties. The increase of such practices creates chains of controllers and processors which are managing personal data on behalf of different entities, which also have the potential to create identity narratives through their interactions.

Beynon-Davies (2011) suggests that the *GOV.UK Verify* system developed by the Government Digital Service (2020a) seems to create:

> [A] private market of Identity Providers with the aspiration that consumers would be able to choose which entity they trust more to handle their identification. It would also allow users to manage multiple electronic identities, having different accounts with separate providers. This way a user can choose where to deploy each identity and for which use (Beynon-Davies, 2011, p. 57)

These issues are highly relevant for this case study but need some further critical reflection, for example, regarding the commercialisation of identities, and the question of the agency of users which often seems to be taken for granted.

In the following section, key legislation which has impacted upon this growth of digital identities will be discussed.

6.2.4 Main legislation and practices

The EU and UK legislation played an important role in shaping the debate on digital identities. As the need for digital identities grows, regulations are being introduced for controlling and standardising existing practices, and for shaping future directions of specific sectors (such as Marketing and Finance). The need to have mechanisms for the identification and authentication of identities via secure systems (across various industries and countries) led to the creation of important legislation by the EU, such as the Anti-Money Laundering legislation, resulting from two EU Directives (see a below); the Open Banking system (Brodsky & Oakes, 2017); the Know your Customer/Client (KYC) process (Arner et al., 2016; Ruce, 2011); the eIDAS Regulation (Regulation No 910/2014 of the European Parliament and of the Council of 23 July 2014 on Electronic Identification and Trust Services for Electronic Transactions in the Internal Market and Repealing Directive 1999/93/EC (EIDAS Regulation), 2014).

a. The Anti-Money Laundering legislation was created to strengthen the EU's financial system against money laundering and terrorist financing, forcing organisations to adopt stricter control on identities. The legislation is the result of a series of European Directives, the Fourth Money Laundering Directive (European Parliament and Council, 2015), and the Fifth Money Laundering Directive (European Parliament and Council, 2018) (transposed into the UK in January 2020), which identified risks and imposed fines for organisations failing to correctly identify customers.

b. The Open Banking system includes a series of reforms that promote the use of open-source technology, transparency, and wide interoperability between different subjects. It is based on a "collaborative model in which banking data is shared through APIs between two or more unaffiliated parties to deliver enhanced capabilities to the marketplace" (Brodsky & Oakes, 2017, p. 2). The Open Banking Working Group noted how Open Banking Standard may need a new approach to identifiers and permit "the identification of parties, resources, devices, applications and products" (Open Banking Working Group, 2018, p. 32). The EU had a relevant role in creating a pragmatic approach to Open Banking via the updated version of the Payment Services Directive/PSD2 (The European Parliament and Council, 2015a) and the promotion of competition via the UK's Open Banking Standard (Brodsky & Oakes, 2017).

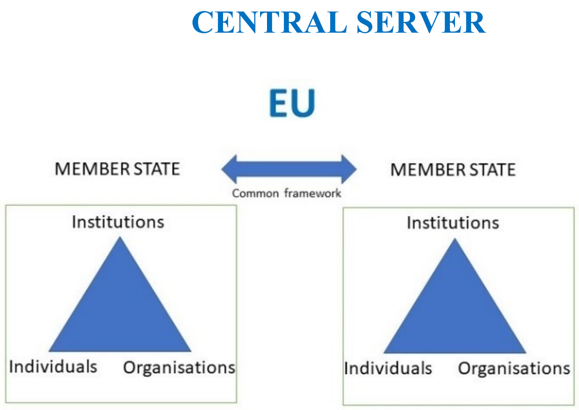
c. The Know your Customer/Client (KYC) guidelines on the identification, suitability and risk of customers are important tools in the management of risk in the financial sector (Arner et al., 2016; Ruce, 2011; The Financial Industry Regulatory Authority, 2011). In marketing, they translated into consumers' profiling practices. Starting with loyalty cards, they have now evolved into more complex systems of ubiquitous surveillance, where different data points linked to offline and online behaviours are merged into digital versions of users.

d. Electronic Identification and Trust Services Regulation for electronic transactions in the internal market /eIDAS Regulation (Regulation No 910/2014 of the European Parliament and of the Council of 23 July 2014 on Electronic Identification and Trust Services for Electronic Transactions in the Internal Market and Repealing Directive 1999/93/EC (EIDAS Regulation), 2014) is one of the most important pieces of legislation on digital identities. Another milestone in the EU regulatory environment aims to create a consistent and interconnected digital identity system able to simplify the bureaucratic complexities within the European Union.

The chosen legislative instrument was once again a Regulation, directly enforceable after two years. The EU created the rules for Electronic Identification (eID) and Electronic Trust Services (eTS), which are key enablers for secure cross-border electronic transactions within the EU and crucial elements of the Digital Single Market (European Commission, 2020b).

The eIDAS Regulation connects citizens, institutions, and organisations.

Figure 6-1 Identity verification via a central EU server



(Design: Chiara Addis)

Electronic Identification (eID) guarantees the identification of a subject who is entitled to receive a service in a Member State. EIDAS guarantees the mutual recognition of Electronic Identification within the EU (Figure 6-1), being “a trusted verification of a client's identity and compliance with Know Your Customer and Anti-Money Laundering requirements” (European Commission, 2020b). The Electronic Trust Services (eTS) certifies electronic signatures and certification services.

As some specifications in the Regulations were to be clarified by the Member States, the UK published the Electronic Identification and Trust Services for Electronic Transactions Regulations 2016 (The Electronic Identification and Trust Services for Electronic Transactions Regulations 2016, 2016), and the UK eIDAS Regulations, partially amended by the Data Protection Act 2018 to include the ICO Commissioner Investigative power (ICO, 2020c). The services set up by eIDAS help “verify the identity of individuals and businesses online or the authenticity of electronic documents” (ibidem).

Of particular interest for this research, the ICO is the “the supervisory body for UK trust service providers…and [it]…can grant and revoke qualified status for trust service providers established in the UK, report on security breaches, carry out audits and take enforcement action" (ibidem).

6.2.5 Service providers

Different countries have managed digital identity requirements in different ways.

Prior to the eIDAS Regulation, the UK had already implemented the *GOV.UK Verify* system developed by the Government Digital Service (Government Digital Service, 2020a). *GOV.UK Verify* was a new Electronic Identity Management (eIDM) system presented as a secure way to prove the identity of a subject online. It can be used by individuals only for services provided online by public authorities, for example, for filing tax, applying for benefits or for a basic Disclosure and Barring Service (DBS) check (Stalla-Bourdillon et al., 2018).

This new eIDM is particularly important. For the first time, the Government can authorise Identity Providers (IdP) to verify identities by checking the public registers, such as National Passports and Driving Licensing Authority (DVLA) (Tsakalakis et al., 2016). The current UK Identity Providers (IdP) certified to verify identities are Barclays, Digidentity, Experian, Post Office, and SecureIdentity (Government Digital Service, 2020b).

National eIDMs, such as *GOV.UK Verify*, can be used when dealing with institutions of EU states, thanks to the interoperability between states created by the eIDAS Regulation.

Differently from the UK, continental European countries have created centralised identity systems. Some governments, like Germany, manage their own Electronic Identity Management system (Tsakalakis et al., 2016) validating identities against official identity records. Similar schemes exist in Italy and Spain.

Other systems act as Identity Providers, like the one being used in Estonia, which created a unique database using an open-source software used by both institutions and businesses. The same technology, which centralised the digital identity of the population, is also used in Finland, Iceland, and Japan (Huber, 2019).

Some projects aimed at the digital identity market are also being developed in continental Europe. For example, the project Irma (Privacy by Design Foundation, 2020), aims at validating identities by gathering reliable data collected by the users, revealing only relevant attributes for specific situations, creating a decentralised architecture (permitting storage on users 'phones), and making the system free and open source.

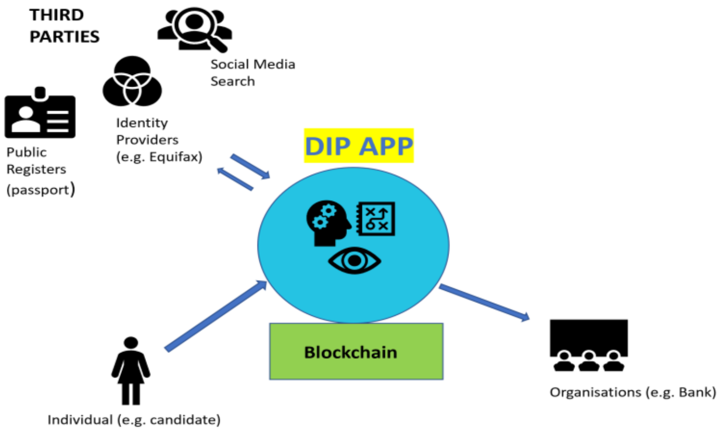
In the following, the project which forms the core of the case study will be described.

6.3 The Digital Identity Project (DIP)

The Digital Identity Project/DIP (not the real name as per confidentiality requirements) is being developed by a new UK start-up that is creating a digital identity solution aimed at individuals and organisations operating in different sectors. The start-up was created by a group of professionals working in the financial sector after realising there was the need for an identity system able to reduce the burden of identity requirements. The project aims to facilitate the exchange of information between businesses and individuals, reducing the cost of compliance, and providing a data trust for individuals who want to use the system as a data wallet for their identity documents. The product is being developed with consideration of different sectors and stakeholders. For example, the financial sector (e.g., banks), housing (e.g., estate agents/landlords), construction, charities. It is expected to be used for various services and for checking the identity and service entitlement of various subjects, e.g., the identity and right to work of a potential employee, the right to rent of a potential tenant, or vetting the admission to or entitlement to a membership of a club. The service is offered via an app, which collects and verifies the authenticity of the documents provided by the subject. Three levels of identity assurance are provided (unverified, partially verified, and verified).

The system created by the DIP project allows the exchange of data between different stakeholders (Figure 6-2).

Figure 6-2 Different stakeholders processing data


THIRD PARTIES
Social Media Search
Identity Providers (e.g. Equifax)
Public Registers (passport)
DIP APP
Blockchain
Individual (e.g. candidate)
Organisations (e.g. Bank)

(Design: Chiara Addis)

Documents are provided by individuals, scanned, and checked for authenticity using different services provided by third parties. DIP then verifies the identity of the person. The App is based on a centralised model (data is stored and managed by the organisation), and documents can then be kept in the digital wallet and used by the individual for other projects, for example for community projects.

Individuals can use the app to create multiple personas to be used with different organisations. The authentication is done via different means, according to different levels of compliance required for specific sectors and circumstances. Chips in passports can be scanned and face recognition is used to verify the authenticity of the pictures on passports and driving licences. A Social Media Search can also be provided, as digital footprints can be used to inform the decision. A certificate attesting the truthful identity and the entitlement to a service are provided to both organisations and subjects. Changes affecting data, e.g., in cases of expired documents, trigger alerts to the parties. Exchanged personal data, such as documents and ID, are kept on Blockchain. Data is stored in the UK (Microsoft Azure Cloud). Compliance and DP are considered as important assets for the product, which is presented as an important tool to reduce the risk of data loss.

6.4 Analysis

The case study was developed using semi-structured interviews, document analysis (two DPIAs, privacy policy, promotional material, website), and participant observations (more details in Appendix B). In this section, the analysis and key themes identified in the data are presented. Their specific sources are indicated by a code (e.g., CS2/1 = interview with participant 1, CS2/D = document, CS2/O = observations). The information is presented considering the three main areas (AI, GDPR and FAT). More detailed analyses of documents and observations can be found in Appendix B.

6.4.1 AI management

a. Strategic decisions

The digital identity market is a growing one. The DIP project was created after realising there was an increasing need for correctly identifying high numbers of individuals (often potential customers or job applicants) firstly in the banking sector and then in other industries. The project is presented as personal data exchange, a solution able to solve organisations’ and individuals’ different needs, reducing compliance costs, increasing privacy, transparency and control on data, and inspiring social change. The Data Trust is also presented as an answer to the government request to improve access to open data to encourage the growth of the AI industry in the UK (CS2/D). The technology developed by DIP supports, confirms, adds some assurance, and reduces the need to '*humanise*' (CS2/2). Furthermore, some of those practices were considered unethical, as similar checks are often performed by underpaid and exploited workers based in other countries (CS2/2). Additionally, the political situation created by Brexit increases the need for the project, as the checks conducted on the rights of people to access services after the UK has left the EU are more stringent. However, the highly charged political climate was also perceived as a risk, as *'you don't know how that technology could be misused in the public domain'* (CS2/1).

The app aims at customers in different locations. It was planned to be multilingual and multijurisdictional in order to be adaptable to other countries. Brexit created uncertainty, and the internationalisation of the project is now more difficult and challenging. DIP was considering the possibility of opening an office in Belfast, depending on how the situation was going to develop in the following 12 months.

Even though the Identity market is growing, DIP did not see other competitors developing similar projects, as they are developing a new idea and a new product. Having a great idea for an app on its own is not considered enough (CS2/2), as that needs to be adapted to different contexts. The company also reported the need to be part of a system, and to have some support from local and national governments.

DIP was interacting with various potential customers. Dealing with small to medium enterprises was more challenging than expected, and this was mainly due to a more laissez-faire attitude to compliance. They were planning to change their strategy and to interact more with their Boards, which was expected to be difficult.

b. The identification process

The app offers different levels of validation and different ways to access the services (e.g., automated or augmented with a human validating the identity). In general, the process consists of different steps: a collection of various personal data from an individual; request of consent; checking according to different levels of validation. Individuals take a selfie, scan, and upload documents with addresses (e.g., bills) and photo IDs (e.g., driving licenses and passports). The selfie is also used to validate consent. Passports 'NFC chips are read by the app, which provides high reliability. '*This is a very good authentication process. Typically, you could forge your passport, but you can't forge the chip'* (CS2/2).

DIP is not connected to Public Registers. They cannot certify the validity of a document, but they can certify that driving licences and passports are not on the lists of stolen documents. '*We can check it to the best of our abilities'* (CS2/1), which shows the best intention, but they are aware of the limits of their means. Email and phone numbers are also collected. Various checks are performed via third parties, including checks on the electoral register, sanction lists, facial recognition and address recognition using ML systems.

Facial recognition checks are provided by Microsoft Azure Cognitive Services. The picture's data is stored on Blockchain and not stored on the device, considered unsafe in case it is stolen. Address recognition is provided both by a company specialising in document recognition and by Experian or Equifax (which confirm if the address is on their databases).

Adopting a triangulation approach, the information in the ID document is verified via Data Brokers (Credit Reference Agencies) and social media and adverse media checks.

Data from various sources is merged to compile the digital identity of the individual. When checks are passed, and the identity is confirmed, a certificate is provided to both the candidate and the organisation.

The meaning of identity, its complexity, the importance of the different elements that make up an identity and the connection to a group all emerged during the interview with CS2/1 who clarified that they did not think passports and driving licences are enough to prove someone's identity.

> *We do not think identity is your passport or your driving licence, that's not relevant to us…it says nothing about it, it gives you a number, but it doesn't say anything about who you are, you know, whom you identify with as a person (CS2/1)*

The participants consider identities as a wider concept, where self-identification and group connections have a significant role in the process of validation.

> *…you've got a group of people who know you by name, and you are known at that particular address, it doesn't have to be your real name, right? The fact that you*

> *are known at that address by a community of people, that means you can be found (CS2/1)*

Furthermore, the app allows users to validate their identities, and also to use different personas while interacting with different subjects, '*it's that broader context, and in that context, one person can have many identities, so you don't have one identity. Do you have many identities? That's okay. We cater for that'* (CS2/1).

c. The quality check processes

A process ending with a positive result leads to the validation of the document (and the identification of the individual). A negative result can have different outcomes, with some yet to be clarified. Potential issues occurring during the validation process are dealt with in different ways. The organisation that is onboarding a candidate is expected to define the policy for when identities are not validated due to an issue with documents. In similar cases, DIP suggested an augmented process. When the verification done by the automated process is not successful (such as facial recognition or document recognition), DIP recommends a human to perform a manual identification of the document. This was expected to be a minor issue as the app can perform both authentications, '*self-service or aided self-service'* (CS2/2) via a partial technology-driven process. For example, visiting the branch of a bank and having the documents checked would solve the temporary issue and complete the identification process.

When the automated process cannot be completed, this was usually seen as the result of a human error. For example, an issue while uploading the documents.

> *The algorithms are pretty good…. The challenge for facial recognition really is in the one to many* […] *you have a database of lots of criminal faces and you want to try and find me in that, as one face against many* […] *there is facial recognition, one to many, and then there is facial recognition one to one…the accuracy level is quite high… (CS2/2)*

A poor-quality selfie taken with a low-quality camera, poor light in the room or from the wrong picture's angle can all impact the quality of the picture, affecting the check.

> *Is it the technology that's broken or is it the fact that I've chosen to do this? Is it a bug in the software or is it actually the user?* […] *We will use the technology to*

> *the best of our ability to capture a good image, to make sure the data is accurate (CS2/2)*

The same participant then clarified this point, '*the technology is not going to be bad. It could be the people that use it'* (CS2/2).

The results of the checks can produce unexpected outcomes and dealing with them can cause a dilemma for organisations. CS2/1 recalled the case of a Social Media Search that linked a work email of someone to an adult website. That finding was interpreted as resulting from hacking and misuse of the email address. However, it also created in the participant a certain amount of doubt about this explanation:

> *I mean, do we really care? I don't know, but you know…it throws up those types of issues. Yeah. Is that a good thing or a bad thing? I don't know. It depends on….It depends on what the job role is* (CS2/1)

The person was not aware the email had been hacked, but '*…You know, the data is in the public domain, right?*' (CS2/1). If the results on the certificates are not correct, the individual is provided with some contact details of the third parties who performed the checks, to whom the requests for rectifications must be addressed by the individual. CS2/1 noted that this can take a long time. By paying a fee, individuals can:

> *[G]o and maintain their KY*C [Know Your Customer] *record. If that record is incorrect for some reason…then it's actually the people who've provided the data to Experian* [credit scoring company] *they have to go back to. So, if you've got a fraud mark again, and that's showing up on your credit report, but if that's come from Barclays bank, you have to go to Barclays Bank to have it corrected there, and then all the way through* […] *it can be a difficult process* (CS2/1)

The possibility of correcting the mistakes directly with the onboarding organisation is currently being taken into consideration. DIP is considering creating a feature in the app which would permit direct communication between the organisation and the candidate, '*the applicant gets the chance to challenge the content because there is a chance that the data is not right'* (CS2/1).

Therefore, individuals can have their data rectified by contacting the third parties that provided DIP with the results, and they might be able to correct the mistakes directly with the organisation. The possibility to challenge the content directly with DIP was neither planned nor discussed.

No particular concerns were mentioned about the third parties' modus operandi, for example, with regards to their methods of sourcing and processing data, or about the quality of the results received.

d. Perception of risks

AI was perceived as a neutral technology. The risks linked to AI were not perceived as linked to the technology per se (not inherently bad), but to the way it is used by a company. Overestimating what AI can achieve, and the confusion between AI and ML in the general debate were both discussed by CS2/2. People were said to be expecting too much from a technology assumed to solve all the problems. The name AI was also thought to increase the expectations, as people think of machines using intelligence like humans, while in reality ML is very good in being applied to specific situations like fraud prevention (CS2/2). Specific risks identified in autonomous decisions made by AI systems were seen as a real source of concern and a potential threat to society (CS2/1). This is particularly the case with decisions made in areas the general public find difficult to understand. Yet, autonomous decisions within the app are perceived to be less risky. The more autonomous the process is, the more secure it is perceived to be. The automated process (self-service) is seen as risk-free, as explained by CS2/2: '*the more control you give to the customer, you increase the chance of that journey going wrong* [...] *and a fraudster will do it on purpose* [...] *will intentionally try to become an exception*'.

A further and different risk was identified by DIP while looking for organisations interested in using the app. The app seemed to raise some expectations regarding the internal capacity of the organisations, which can make visible the need for a transformation process: '*The app is a very successful consultancy tool because it raises questions inside an organisation* [...] *it actually highlights that there were areas of weakness within the recruitment process that need to actually be addressed* '(CS2/1). This has the power to disrupt the organisation significantly more than adopting an app, and not every organisation is ready to admit the need, or can afford, to start a transformation project.

This can become a risk for DIP, which is currently focused on seeking business partners.

e. Summary

Under AI management, some key elements were identified. For example,

- Decision-making in response to external factors and personal ethics.
- External and internal processes.
- Role of external stakeholders.
- Differences in the understanding of risks, technology, and the role of humans.

Those elements will inform the model in Chapter 8.

6.4.2 GDPR

a. Top-down approach vs self-regulation

GDPR was reported as *'the core of everything'* (CS2/1) DIP does. The Regulation is fundamental in increasing individuals' empowerment and in promoting the company with customers and business partners. Nevertheless, CS2/1 believed the protection provided by the Regulation was insufficient, as '*…the areas that are of interest to us are the areas that it doesn't cover'*. Considered to be a top-down regulation designed to deal with specific cases (such as the right to be Forgotten), GDPR was not considered as able to provide proper protection around some of the issues that, for example, led to the Windrush scandal (Gentleman, 2019), where individuals could not prove their citizenship rights.

The complexity of the privacy policies, lack of enforcement and a lack of individuals' agency and control were some of the problems identified within the realm of DP. The GDPR was seen as incapable of remediating due to a lack of control around data and data sharing after the individuals have given their consent.

> *…once you give your consent to an organisation…it's gone. We all know that GDPR is supposed to be protecting that, but the reality is nobody has time to read all the privacy policies you consent to, you sign to…and even if you could read them, you can't understand them…so, it's a question of whether that gives them informed consent or not* (CS2/1)

The top-down regulatory environments were not considered to work in these contexts, and the result is a tick box compliance that does not achieve the necessary cultural change. Technology comes first, and it changes the culture: '*Mobiles have changed the culture in* *about 10 years without any regulation* […] *Yeah, so you've got analogue regulators in a digital world'* (CS2/1). DIP would prefer a bottom-up approach to regulations where communities of practice can develop their own standards.

b. GDPR requirements and stakeholders

GDPR had also proved significant in the search for commercial partners. The GDPR creates an obligation for organisations handling personal data to register with the ICO. CS2/1 recalls the case of an organisation interested in using the app to verify potential customers. The negotiation ended abruptly when the company was found not compliant. The organisation was not registered and had no intention of doing it. Even though they were aware of potential fines, they did not believe they were at risk of hacking, the only reason considered for registering. The hostility against the registration was something DIP had not anticipated, and that was particularly important as DIP was trying to work actively with the ICO. They had sent a grant application and they were hoping to become part of the ICO Sandbox (ICO, 2020b). That would allow DIP's project to be recognised as a pioneer, and to be researched and supported.

c. Processes around technical and organisational measures

While the exact location of data within organisations can be an issue for many organisations such as banks (where the participants had worked), this does not seem to be the case for DIP.

Data location and storage were not considered an issue, as they knew where their data was (CS2/1). The use of Blockchain gave DIP a high level of confidence. The adoption of a '*hypercube database'* was also mentioned as a further security measure taken to protect data and make it hard to identify (CS2/2) (potentially pseudo-anonymising data).

Both participants made clear it data that is not sold to third parties, nor transformed through the practice of derivative data, reportedly being used by many companies that change the output data by adding a new element and then use it to develop other products (CS2/2).

Data is usually deleted after two weeks unless individuals decide to use the data trust service offered by DIP while dealing with other organisations.

The new GDPR requirements mandating the need to have a DPO and to perform a DPIA were both discussed. The CEO was the current DPO, and a paralegal specialising in DP was acting as a watching brief for DIP.

A DPIA was performed at the end of 2018. While this was considered to be sufficient at the time, they were aware of the need to review it. A new DPIA was created after the interview and sent to the researcher. The most recent DPIA is a long and detailed document about the data trust service (CS2/D) created due to increased risks. For instance, the use of AI for security, the combination of data from multiple sources, and the use of federated identity assurance services (which link identities across different identity systems).

The need '*to review everything around GDPR'* was also mentioned by CS2/1.

d. Identity checks

Empowering people is very important for DIP. People's ability to control and decide whom to give their personal data to is central within the project. This was mentioned as an important factor in considering the creation of a data trust to be used by individuals while dealing with different organisations. Identity checks performed using social media and adverse media checks are considered another significant element in increasing the awareness of individuals. Third parties are used to check personal data available on the Internet: *'We will build up a data set that gives us, you know, some level of confidence that that person looks like that, lives at that address…' (CS2/1).*

Databases built by data brokers corroborate the individuals' identities: *'…there are databases that you can call…you are well aware of Experian and Equifax and all these others…You are somewhere….' (CS2/2).* Participants felt strongly about that, as the results coming back from a digital footprint search are important for the individual as well as for the onboarding organisation. Personal data available on the Internet is considered to be public domain, and therefore usable: *'We are not taking data that is private, we are taking data that's, you know, that has been put in the public domain* […] *this is open data that's been made available by the individual* […] *So they have consented for it to be used'* (CS2/1). The availability of the content is dependent on the privacy settings chosen by individuals while setting up their social media profiles: *'It depends on how they've done their privacy settings on the social media platform and how they identified themselves in those platforms* […] *but they may be unaware that they have done it. And that is the issue' (CS2/1).* Despise these concerns, that data was to be used.

Various data is collected. For example, ID documents with pictures, pictures/videos, documents from authorities (such as Council Tax), a check with their local authorities if individuals have recently used their services, social media, bank/financial details.

Organisations can require further data according to their specific needs (CS2/D). After verifying the identity via an automated process, a certificate with the results is sent to both the organisation and the applicant. If there are any inaccuracies in the data, the applicant should let the organisation know immediately. The applicant is also provided with the contact details of each of the third parties who have provided the results. Applicants should contact them directly concerning the results provided.

This section on the GDPR showed the importance of the Regulation in the work of the organisation, and also its limitations.

e. Summary

Other key elements were identified under GDPR. For example:

- Stakeholder management.
- Processes for assessing data quality and risks.
- Self-reflexivity.

6.4.3 FAT

This section reports the participants' opinions on the FAT principles.

a. Fairness

Compliance, fairness, transparency, and protection of individuals' rights are clearly stated in the company's mission (CS2/D). The principle of fairness is very important for DIP. The project was developed to empower people and communities. The Cooperative and Quaker traditions and Open Data movement were all a source of inspiration.

They saw their work as part of the Open Data initiative. The government policy was reported to encourage the openness of data people want to share. They see their work as very important in this cultural shift which strengthens individuals. DIP wants to do good and raise individuals '*agency by giving people the power to choose whom to share their data with, giving them control over their data'* (CS2/2). Individuals are central in their vision, so '*the companies have to come to them…*' (CS2/1). Participants had read specific literature on data exploitation and data capitalism. The book by Zuboff (2018) had a big effect on their thinking, giving them a framework for their project. They were aware of biases, unconscious biases, and of the capacity of ML to reinforce stereotypes and poor behaviour (CS2/1), which can produce discriminatory effects upon specific categories (such as women and BAME people) (CS2/2). Compliance and regulations were seen as necessary means to encourage people to do the right thing, achieve positive outcomes, and do social good (CS2/2). Recent political events, such as the Windrush scandal and Brexit, were viewed as examples of the urgency to empower people, and this was also considered necessary to compensate for the unintended consequences of the technology (CS2/1).

Both participants clarified that the data they are processing belongs to the individuals. They had no intention of providing pictures of faces, documents, or any other data to third parties. They believed that only individuals could consent for their data to be shared, and some hypothetical scenarios were mentioned by CS2/2. For example, in the case of:

- Medical research: '*You might say, I'm happy for my data to be shared with this research company that's investigating cancer'* (CS2/2).
- Mental health online chat: As a safeguard measure against trolls, the administrator of the chat could ask new members for proof of identity. They could also provide their preventive consent to the administrator for sharing their data with another person in the case of a crisis. This appeared to be a less than hypothetical scenario: *'this is very real, because …there is a social group we're speaking to that has that problem'* (CS2/2).
- Compensation: The individual could decide to share their data in exchange for an economic compensation *'so you're giving something, you're getting something*' (CS2/2).

The hypothetical possibility of being asked to provide personal data of customers by institutions, for example the Home Office, was viewed as an interesting challenge, as the consent of the individual is paramount for sharing data with others (CS2/2). They knew of a few local authorities which had been pressed by the Home Office to provide personal data.

Participants reported discussing fairness all the time. Sometimes this could become a problem, as they felt they were educating potential partners on issues they had not thought about before (CS2/1).

b. Accountability

Accountability is very important for the project and the relationship with various stakeholders. DIP was trying to make accountability more explicit, especially in relation to data trust. However, accountability was also an issue for the company. This is due to the system of roles and relationships created amongst stakeholders, which characterised their business model.

Understanding which legal framework is right for their business model, in relation to GDPR and data processed by the different parties, was an ongoing process. That was the case even after having consulted a legal expert, the Open Data Institute and the ICO.

As different entities are involved in handling personal data, different roles and responsibilities were identified from the exchange (CS2/1, CS2/2, CS2/D):

1. DIP-Individual: DIP is the data controller in the relationship with the individual (CS2/1). This is both the case while verifying the identity and while acting as data trust. DIP is given data to perform a function, and it does not sell it or give it to anybody else.
2. DIP-Partner Organisations: Both DIP and the organisations are currently considered joint-controllers. While this was the result of legal advice, DIP felt uncomfortable with this setting. If any of the partners breached DP, DIP would be liable. DIP was looking at changing this arrangement and becoming a processor.

Due to these different responsibilities, DIP had created two different legal entities, one dealing with individuals and another with organisations. While this change had provided DIP with a certain level of clarity, they felt the need to have more.

The potential future collaboration with the ICO via the Sandbox provided a certain level of reassurance, as they were hoping to have an appropriate control around that: *'The liability…let's say as far as the ICO is concerned is manageable. We still have a liability to each individual, of course, but you know, this is a difficult area*' (CS2/1).

Participants were aware that accountability is a critical area.

c. Transparency

Transparency was also very important in relation to the individual. The participants believed organisations should be transparent with their customers, and provide an explanation of the reason behind their decisions:

> *...the challenge for those organisations is...you can have ML, but you have to be able to explain how ML came to the conclusions that it did. Because if the customer says, I want a transparent answer, how did it happen? Now, this typically hasn't occurred before (CS2/2)*

The customers have the right to have a transparent answer, and this is something new.

DIP are transparent with individuals, they make it clear that they do not resell their data, and they provide a certificate report to individuals and organisations, '*so the applicant sees totally what we would see'* (CS2/1).

However, to delete the information held by organisations individuals must directly contact the organisations to request further action. DIP '*can provide a list of organisations with which they have enrolled'* (CS2/D).

This section shows that the participants cared deeply about FAT, also acknowledging some challenges.

d. Summary

The key elements identified under the FAT principles that will inform the model are:

- The connection between a fair use of the technology and the positionality of the stakeholders involved in the processes.
- How fairness, accountability and transparency can be impacted by the involvement of specific stakeholders, such as end-users.

These aspects too were used to create the model.

6.5 Discussion

In the following section, the key elements identified in the analysis of the data will be discussed.

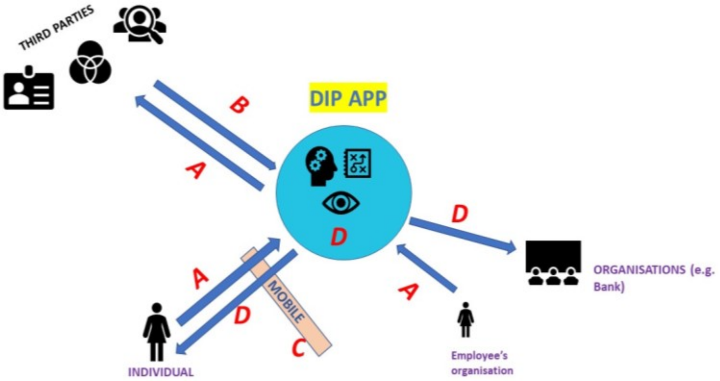
6.5.1 AI management

The experience and knowledge of the participants are remarkable. That is clearly noticeable in this multifaceted project and the work carried out by the company. They are very attentive and careful in planning how to use personal data, and in creating a system that empowers data subjects.

DIP has created a complex business model based on data which includes different stakeholders. Different entities are part of the ecosystem where DIP operates. They process personal data, make judgements and work on the assumptions made on that data (for example, while considering data as personal, non-personal and special category).

- Data is exchanged, but it is also created. The App is presented as a personal data exchange project. However, the findings show a much more complex ecosystem, whose characteristics, and how they are perceived, all impact both on the data and on the internal processes set up by the company. While the exchange of data clearly occurs within the system, personal data is not only exchanged but it is also created by different entities all having a certain degree of responsibility. Data is treated differently according to who is processing it. A clear distinction appears between the flux of data occurring between the DIP App and third parties, and the one between data subjects, App, and organisations. For examples, there are four types of personal data (Figure 6-3) related to one same data subject who is using the App.

Figure 6-3 Types of personal data exchanged between different entities



(Design: Chiara Addis)

a. Data provided by the data subject in order to have their identity verified (e.g., passport).

b. Data resulting from the multimethod search conducted by third parties (e.g., results from social media search or checks with councils).

c. Data produced during the interactions data subjects have with the system via their devices.

d. Data produced by the App after gathering all info from different third parties (certificate).

- Categories of data. DIP clearly considers personal data to be the data provided by the data subjects. Other data does not appear to have this status. For example, the data found in the public domain, the data provided by third parties, or the data created by using their devices (C) (e.g., unique device identifier such as IMEI number or phone number). Such data is not included in the data the subjects can exercise their rights on (e.g., right to access or erasure).

- Facial Recognition and special categories of data. Biometric data are special categories of data clearly defined and protected by the GDPR. According to the DPIA, the App does not process special categories of data. This is not correct. Biometric data are processed in the project. For example, the pictures used for identification purposes processed with Facial Recognition technologies, data concerning health (e.g., medical records), and data related to people's sex life (e.g., potentially resulting or inferred from the Social Media Search). Additionally, the applicant privacy policy clearly prescribes that in the case of suspect criminal activities, data (certificate and biometrics details) can be kept for a period of seven years. This is problematic. Facial Recognition is an extremely powerful and controversial technology that is booming as a means of security and surveillance. Currently banned in various cities for its capacity to be used for political and racial surveillance, the technology is being normalised through its growing use by the public and by private organisations. The type of Facial Recognition used by the project, described by CS2/2 as the '*one to one*', was not considered as risky as other types specified by CS2/2 as '*one to many*'. However, the risks do not only come from it being highly inaccurate, amplifying historical discriminations, but also from the lack of oversight and accountability, especially when used by private organisations.

  The repurpose, retention and erasure of facial recognition images are recurrent concerns with this technology, and while its use by law enforcement is now being more discussed, the same level of scrutiny and accountability in private companies is still far away.

- Assumptions and risks. The validation of the identity of an individual is done considering the data provided by the subject. It is assumed that this data could be wrong or the result of fraud. Therefore, the judgement is suspended until receiving the results of the third-party searches, which provide external validation of the data. Conversely, the data provided by third parties is assumed to be correct, not questioned, and not checked. Moreover, there was no mention of their processes, their sources of data, and their commitment to DP and privacy rights. The possibility of any mistakes resulting from those partnerships was not considered. While the limited time of the interview may have impacted on the number and depth of topics discussed, it is noticeable that no reservations about the work carried out by third parties were identified in either documents or observations. Similarly, the technology and the capacity of algorithms were rarely questioned. AI was in general perceived as a neutral technology, and risks were seen in how other organisations are using AI, but not their companies. Potential issues with the app were often considered the result of improper use or behaviours of individuals. The issues are remediated by the "human in the loop", who manually checks the documents and validates the identities. While some of the risks coming from some external factors (e.g., Brexit) were considered, other risks coming from other factors (e.g., third parties) were not. Overestimating the work done by some stakeholders can have serious consequences. Wrong results can impact on the rights of data subjects (e.g., entitlement to a job or a house) or their reputation (real or wrong data about their sex lives) without the organisation providing the results being aware.

- Judgements. The results received from different third parties are aggregated by DIP, which then issues the certificate to validate the identity and therefore the entitlement to a specific service. Thus, DIP makes a judgement on the data gathered by various parties and based on that it generates another content (certificate), which is itself another personal data.

6.5.2 GDPR

The GDPR is crucial for DIP, for the project, and for finding business partners. Regulation and compliance are seen as necessary means for doing good and increasing the power of individuals. DIP’s focus on compliance had an impact on their relationships with stakeholders. Creating a project aimed at improving compliance has often created some difficulties in finding business partners sharing similar views.

Of interest is how the Regulation is perceived by respondents. They agree on the importance of having some regulation, which is also necessary to counteract the data exploitation engaged in by big companies. However, they see the GDPR as having many flaws:

- Top-down regulation. The GDPR is not considered able to facilitate and manage innovation. A bottom-up approach created by communities of practitioners was considered more effective in providing organisations with both more specialised knowledge and the necessary flexibility to adapt to the pace of the technology.
- The GDPR is not considered appropriate for protecting individuals, who accept complex privacy policies and provide personal data without understanding the implications. Regulators are not enforcing the new rules. Consent is seen as the key factor in enabling data subjects to exercise power and control over their data. By strengthening the consent and empowering people, DIP expects to avoid situations similar to those created by the Windrush scandal. As data subjects are central to the system created by DIP, their consent is instrumental and necessary for the provision of legitimacy to the processing.

Some elements in the DIP’s understanding and perception of the GDPR are particularly interesting and deserve to be looked at with attention. The type of regulation hoped for, a self-regulated one, could perhaps provide more involvement for practitioners and experts, but this could potentially be less effective in protecting personal data. A self-regulated system, specific to an industry, would not be as effective as a top-down one in capturing and regulating organisations operating in different sectors. The use and exploitation of personal data are often carried out by companies that are part of a complex ecosystem. Various actors operate data exchanges, such as start-ups, big data brokers or social networks (like Equifax or Facebook), and this exchange often occurs internationally via cross-border exchanges. The protection offered by bottom-up regulations might prove to be extremely insufficient for protecting personal data against highly influential geopolitical actors accessing data to influence the public discourse.

DIP is creating a system where individuals are at the centre. Similar projects which strengthen the roles of people and communities are certainly positive and very much needed.

The focus of this project on the management of personal data via consent can increase agency and transparency. However, this can happen only on part of the process.

Focusing this project on consent, as the answer to a lack of power and control, has the potential to be limiting and counterproductive. There are several reasons:

- The system presumes data subjects to be informed and aware when providing their consent. The organisations requesting data are expected to inform the people, who are expected to understand all the implications. Yet, data subjects can still provide uninformed consent, therefore incurring the same drawbacks as other systems.
- The data subject is an element of a more complex system. Other entities can potentially and maliciously influence the decisions made by the individuals.
- The use of data found in the public domain is complex. The information available online is considered public domain if accessible by a member of the general public (without specialised knowledge or research skills), and if it is personal data, it is still subject to the GDPR. The research carried out in the public domain is important, for example, for conducting fraud investigations via Open-source intelligence gathering and analysis (OSINT) techniques. Such practices are often used by organisations to manage online reputation. The social media search carried out by DIP is an example of OSINT. DIP uses third parties to gather information about individuals who, by providing their consent, give further legitimacy to a search that is already intended as lawful (being in the public domain).

The understanding of this process by DIP is very interesting. According to DIP, the data found in the public domain is no longer seen as private data, but as open data made available by individuals when choosing a visible profile. Consent is again relied upon to justify the processing of data, and this happens at two different moments:

1. Consent is provided to a social media platform by choosing a visible profile (Public-domain). This choice allegedly impacts on the nature of the personal data, turning it from personal into open data and therefore usable. However, consent cannot change the nature of the data. Data in the public domain is still personal data and therefore still subject to DP legislation. When publicly available data is used for profiling, data subjects should still be made aware that their data is being processed, and of the consequences of such profiling, and be given the possibility to object to it. The recent

case of an app that gathered billions of facial images of unaware people (from various social media) to create facial recognition databases has caused many platforms to send cease-and-desist letters for the breach of their privacy policies which prohibit data scraping (Sobel, 2020). As pictures are considered biometric data when they are collected and processed to identify a person, such collection without the explicit permission of the data subjects is unlawful per the GDPR. Moreover, data subjects should also be informed with a privacy policy about the existence of profiling and its consequences. Therefore, the use of information in the public domain is not free from the GDPR requirements.

2. Consent is provided to DIP to perform Social Media Searches via their Third Parties. That provides DIP with a legitimate basis for the search in the public domain. DIP then uses this basis to authorise third parties to perform the search on its behalf.

Individuals are often unaware of their digital footprints, and this can be due to two different situations: individuals do not know they have selected a public profile and/or do not know about the data brokers creating databases of personal data. By merging content from different sources, they profile individuals ’current behaviour and make predictions on future behaviour, creating a huge amount of content. Such practices raise many questions in relation to their lawfulness, the effect they can have on people’s lives, and the effect on the management of information within organisations. The dilemma mentioned by CS2/1 resulting from finding a work email linked to an adult website is an example of the latter. The reliability of the information could not be verified, leaving the manager with a potentially very difficult choice.

This over-reliance on consent can risk becoming a search for meaningless consent (Edwards & Veale, 2017). Continuously focusing on consent seems to imply the existence of only one lawful basis for processing personal data. But the GDPR states that consent is one of the lawful bases (Art 6). Others legitimate reasons for processing personal data are, for example, situations where processing is necessary for the performance of a contract, or compliance with a legal obligation.

Leaving the specific considerations on consent, imbalance of power and employment aside for a moment, if people want a job they have to sign a contract and they have to provide their personal data. Similarly, potential customers have the legal obligation to provide their personal data to the banks, which then have the legal obligation to check their identities.

In similar cases, giving consent, and providing data, succeed a pre-existing obligation that justifies the processing. Then, consent is not the only lawful basis as other bases make processing necessary and lawful, and the power to give consent is not free but is conditioned by other legal requirements. Therefore, aiming to strengthen an element of the system (the data subject) by focusing only on one lawful basis (consent) can be less effective than expected.

Increasing consent cannot be sufficient to compensate for the lack of agency already existing in a system where other actors, much more powerful, create content that contributes to the online identities of individuals. Even if the search is done instrumentally to increase individuals ’awareness, the results cannot compensate for the existing gap between the increased power from knowing the information, and the power of data brokers to continuously create online identities.

Consequently, individuals ’influence and agency might be overestimated. The researcher is aware that the project does not have the capacity to resolve all the flaws of the law and the issues of agency. However, a wider and holistic approach inclusive of other factors and actors would greatly benefit this project so focused on empowering people.

6.5.3 FAT

a. Fairness and transparency

Doing good, increasing awareness and power of individuals, and contributing to social change are all important factors for DIP. Their desire to use the project to create a new business model able to facilitate the empowerment of individuals to control their own data, while also facilitating social change, appeared to be a strong motivational factor (CS2/O). Compliance is considered instrumental for doing good, and for reducing the risk of inadequate security measures. Empowerment comes with greater transparency. The visibility of the exchange of data and the possibility to influence it with consent, provide more control to data subjects. DIP’s experiences and collective work history provided a deep understanding of HR management, IT Change and Transformation, and Counter Fraud and Security management practices. Their backgrounds and interests (e.g., Cooperative and Quaker traditions) had a major influence in shaping their innovative approach aimed at increasing fairness for those using their app. The inclusion of some ML components amongst the app’s technologies urged them to become more informed about AI. They consider AI a risk for society and can see the harm AI could cause when used by other companies. They advocate for transparency and believe other companies should provide explanations for their decisions. They were more aware of the myths surrounding AI, and they knew about the capacity of ML to reproduce past discriminations.

The app was seen as an important instrument to strengthen compliance and provide individuals with a means to prove their rights and their entitlement to services. However, some elements which could influence or create unfair outcomes did not emerge, or went unnoticed. DIP is not selling customer data, and they are very clear about the unfairness of this practice. Yet, they are receiving data that is provided by their third parties. They are data brokers trading data. The Data Broker industry is currently being questioned by activists and institutions and regulatory bodies (i.e., EU and ICO) for unlawful, unfair, and pervasive practices, poor technical efficiency, and their use of ML algorithms that is far from being perfect. All those elements contribute to creating personal data whose accuracy, and legitimacy, are open to question.

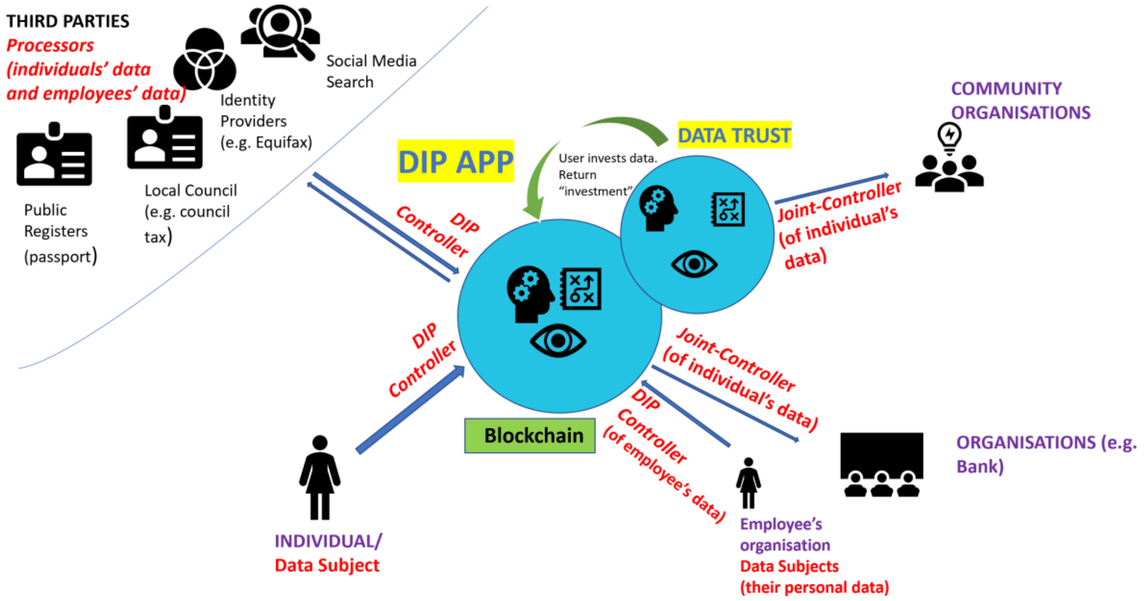
This is highly problematic. The data provided by data brokers informs the customers 'digital identities, which are now more important than civic ones in proving the entitlement to individuals 'rights. Yet many issues were not discussed in the project, such as the role of third parties, their practices, and the possibility of receiving biased data or data resulting from unfair practices. Similarly, while some controls had been considered on certain data stored in the Data Trust (e.g., expired ID documents), no specific measures were planned regarding quality checks or periodic controls on the data received from data brokers.

Furthermore, the researcher noticed a peculiar phenomenon. DIP showed a certain level of self-reflexivity. For example, they reported discussing if they should be doing the project at all. Having the time and space for discussing possible doubts on projects using AI is certainly very positive, much welcome, and a sign of the organisation's maturity. They also displayed a high degree of awareness about dangerous practices adopted by other companies: a rare level of awareness amongst start-ups. However, while they are focused on the protection of personal data, they do not seem to be fully aware of the impact their practices can have on the identity validation process. A discrepancy seems to exist between the effective role played by DIP and how they perceived it. DIP seems to underestimate their role and their power in influencing potentially unfair practices. But at the same time, they seem to overestimate the power of the app in creating good outcomes for data subjects.

b. Accountability

This case raises an important issue in relation to the accountability, responsibility and knowledge that different entities processing personal data have. Accountability is crucial for DIP and its business model. Understanding which legal framework is right for their business model in relation to GDPR was an ongoing process. Making accountability more explicit is one of the objectives of the project. Yet it is also a critical one, due to their particular business model that includes different stakeholders processing personal data. The project is a good example of a growing industry, and it is particularly interesting and useful in better understanding the role of data brokers, organisations, data subjects and their responsibility according to the GDPR. Similar projects created chains of controllers and processors which have the potential to create narratives of identities and different degree of accountability through their interactions.

Figure 6-4 Complexity of the ecosystem where DIP operates



(Design: Chiara Addis)

The analysis of the data revealed that:

- DIP processes data of two different data subjects: applicants and employees of the organisations managing data.
- DIP is the controller of the data subjects' data.
- Third parties are processors.

- DIP is a joint controller with organisations using the data.

The joint controllership is a source of concern for DIP. They do not believe they should be held responsible for something other organisations are doing, and they are hoping to become mere processors. This raises an interesting point. Becoming a processor would be rather difficult considering the high decisional degree of DIP. DIP is the centre of the complex system and makes important decisions on data. Yet joint controllership is also not completely adequate. According to the GDPR, joint controllers jointly determine the purposes and means of processing. That does not seem to be the case in this project, as two controllers share the same personal data for purposes that are not exactly the same for both. This situation appears to suggest a more "separated controllership" relation, similar to the one recalled by the European Data Protection Supervisor (European Data Protection Supervisor, 2019, p. 10). DIP and other organisations (e.g., banks, landlords, community groups) do not jointly determine the purpose of the processing. Organisations process personal data for a specific purpose, in this case compliance, while DIP processes data to check if those documents are not the result of fraud. Additionally, each party processes data by using means that are independent of those used by the other party. Consequently, it is highly likely that DIP is an independent controller and therefore not responsible for the processing done by other organisations. This should solve DIP's concerns about being liable for improper use of data by other parties. Yet, this does not exonerate DIP from its obligations as controller. That is linked to another important element identified in the research. The project does not provide a strong process for rectifying mistakes or errors found in the data. When the App fails to upload the documents, the issue is usually seen as resulting from improper use. The App and its technologies (such as ML) remain unquestioned. When the results of the checks performed by the App are negative, and the identity is not validated, individuals are given the details of the third parties (to contact in order to correct the error), or of the hiring organisation (which in principle could disregard the request).

The method of ensuring accountability could be improved. The current process is not strong enough and this weakens the impacts on DIP's role and obligations as controller.

DIP is at the centre of the system, and it is not a mere exchanger of data. It makes many choices (e.g., which third parties, what kinds of checks), aggregates the data received from the processors, make a judgement on that data, and creates a certificate (personal data).

Therefore, DIP has power and responsibilities which should be conveyed into a process enabling DIP to respond directly to data subjects.

6.5.4 CS2 Key elements

The key elements of the praxis of CS2 are identified in the table below:

Table 6-1 Key elements of the CS2 praxis

Technology	• Technology considered neutral and efficient • No awareness of potential errors in data and technology • Human actions considered prone to mistakes (end-user) vs human actions (staff) capable of correcting errors in processes • ML and different understanding - self-reflexivity
People	• Empowering people/end-users by using an ecosystem that exploits personal data • Relationship between people and praxis of technology • Different understanding and expectations • Data Subjects usually unaware of personal data trade
Processes	• Fewer internal processes than a traditional organisation BUT more external processes with various entities. Increased risks • Overestimating data quality
Stakeholders	• A complex external ecosystem of third parties/stakeholders providing specific services • Data brokers merging information
Decision-making	• Decisions around strategy, business model, data, and relationships with stakeholders

Power	• Positionality of stakeholders • Impact on end-users can be massive
Innovation	• Holistic approach considering internal and external contexts

6.6 Conclusion

This chapter presented the work done by DIP, an organisation developing a project around digital identity. It started introducing the main elements in the discourse around identities, such as the common meaning of identity, the differences between the UK and other traditions, and the main service providers. After illustrating the main characteristics of the project, the chapter presented the analysis of the interviews, documents, and observations. with consideration of the understanding and practices around AI, GDPR, and FAT.

The chapter then ended with a comprehensive discussion of the key elements emerging from the analysis. This revealed the unconcerned use of ML within the project, the central importance of the GDPR but also its perceived limits, and a complex picture around stakeholder management and multiparty responsibility.

CHAPTER 7: THEORETICAL FRAMEWORK – CRITICAL AI&DP MANAGEMENT (*CRAIDA*)

7.1 Introduction

This chapter presents the key analytical assumption within critical AI&DP/*CRAIDA*, a new theoretical framework developed to critically theorize the complex relationships between AI, DP, and FAT within organisational contexts.

Firstly, it presents the two theoretical traditions informing a new synthesised theoretical approach: responsible research and innovation (RRI) and critical theory of technology (CTT), the latter with a specific focus on the work carried out by Andrew Feenberg. Secondly, it describes the new theoretical framework developed for this dissertation, and the main components of which were chosen to analyse organisational contexts. Thirdly, the chapter presents a detailed analysis of the research context carried out using the critical AI&DP Management framework, with a final explanation for using critical theory at a late stage. In the following section, responsible research and innovation (RRI) will be presented.

7.2 The responsible research and innovation framework and the key elements

RRI is an EU governance framework for Research and Innovation, and it is a key action of Horizon 2020, the financial instrument implementing EU research and innovation policy (European Commission, 2020a). The framework has been adopted by major research funders including the European Commission and the UK Engineering and Physical Science Research Council (EPSRC), and it is also used for developing Quantum technologies (Inglesant et al., 2018).

The framework is:

> a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products (in order to allow a proper embedding of scientific and technological advances in our society) (van den Hoven, 2013, p. 63)

RRI aims to improve Research and Innovation (R&I) by satisfying various societal challenges and needs, empowering different actors, and providing a more holistic and collective approach.

Particularly important in educating researchers and practitioners around key guiding principles and best practice has been the work carried out by ORBIT (Orbit, 2020), B. C. Stahl (2012, 2013; 2018), B. C. Stahl et al. (2017), and the EU project RRI Tools community (RRI Tools, 2020).

RRI is based on four principles (Orbit, 2020) (fig 7.1): Process (speed of innovation and diffusion, which includes all R&I activities); Product (inclusive of outcomes and effects of use); Purpose (inclusive of the reason, ethics, acceptability, and desirability); People (inclusive of questions about the correct choice of stakeholders, user engagement, e.g., who decides and benefits).

The societal challenges are included into seven categories (RRI Tools, 2020):

a. Health, demographic change, and wellbeing.

b. Food security and sustainable agriculture, forestry, marine and bioeconomy.

c. Climate action, environment, and resources.

d. Smart, green, and integrated transport.

e. Climate action. Secure, clean, and efficient energy.

f. Europe in a changing world: inclusive, innovative, and reflective societies.

g. Secure society, freedom, and security of Europe and its citizens.

R&I has often been seen as lacking effective stakeholder engagement and genuine consideration for societal values (Owen, 2014; B. C. Stahl et al., 2017). RRI aims to fill these gaps by:

a. Including key actors:

    - Policymakers (inclusion at different levels, comprehensive of organisations' executives, regulators, and funding agencies).

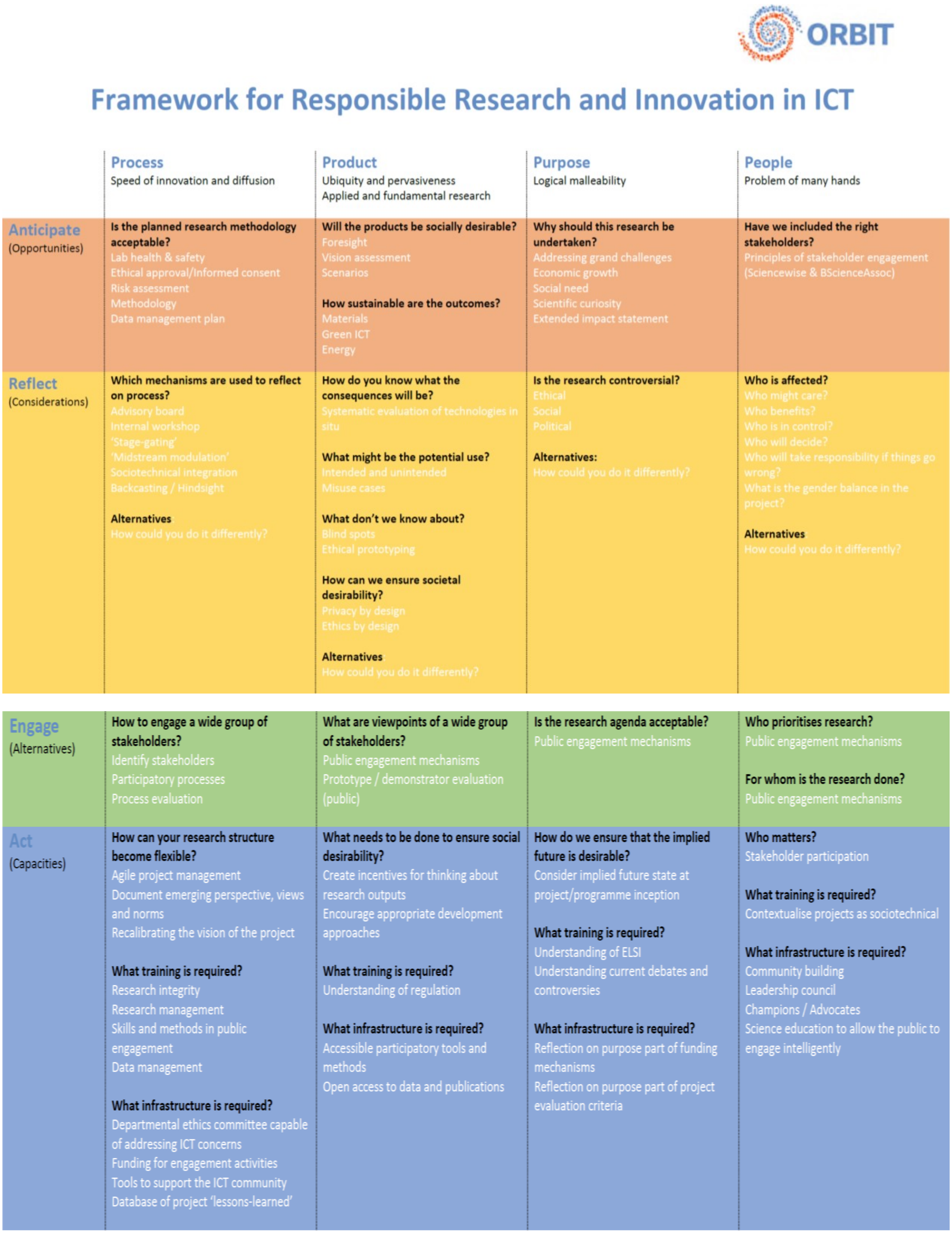
    - Research and innovation communities (all those involved formally and informally).

- Education community (from researchers to students, comprehensive of people involved formally and informally at different levels).
- Business and industry (from professionals to corporations, in-house and outsourced consultants involved with innovation).
- Civil society organisations (e.g., community groups, foundations, charities).

b. Considering key societal values, identified as:

- Ethics (inclusive of research integrity and ethical acceptability of the outcomes).
- Gender equality (resources, bodies, and research dimension).

Figure 7-1 Framework for RRI and innovation in ICT



ORBIT

Framework for Responsible Research and Innovation in ICT

	Process Speed of innovation and diffusion	**Product** Ubiquity and pervasiveness Applied and fundamental research	**Purpose** Logical malleability	**People** Problem of many hands
Anticipate (Opportunities)	**Is the planned research methodology acceptable?** Lab health & safety Ethical approval/Informed consent Risk assessment Methodology Data management plan	**Will the products be socially desirable?** Foresight Vision assessment Scenarios **How sustainable are the outcomes?** Materials Green ICT Energy	**Why should this research be undertaken?** Addressing grand challenges Economic growth Social need Scientific curiosity Extended impact statement	**Have we included the right stakeholders?** Principles of stakeholder engagement (Sciencewise & BScienceAssoc)
Reflect (Considerations)	**Which mechanisms are used to reflect on process?** Advisory board Internal workshop 'Stage-gating' 'Midstream modulation' Sociotechnical integration Backcasting / Hindsight **Alternatives** How could you do it differently?	**How do you know what the consequences will be?** Systematic evaluation of technologies in situ **What might be the potential use?** Intended and unintended Misuse cases **What don't we know about?** Blind spots Ethical prototyping **How can we ensure societal desirability?** Privacy by design Ethics by design **Alternatives** How could you do it differently?	**Is the research controversial?** Ethical Social Political **Alternatives:** How could you do it differently?	**Who is affected?** Who might care? Who benefits? Who is in control? Who will decide? Who will take responsibility if things go wrong? What is the gender balance in the project? **Alternatives** How could you do it differently?
Engage (Alternatives)	**How to engage a wide group of stakeholders?** Identify stakeholders Participatory processes Process evaluation	**What are viewpoints of a wide group of stakeholders?** Public engagement mechanisms Prototype / demonstrator evaluation (public)	**Is the research agenda acceptable?** Public engagement mechanisms	**Who prioritises research?** Public engagement mechanisms **For whom is the research done?** Public engagement mechanisms
Act (Capacities)	**How can your research structure become flexible?** Agile project management Document emerging perspective, views and norms Recalibrating the vision of the project **What training is required?** Research integrity Research management Skills and methods in public engagement Data management **What infrastructure is required?** Departmental ethics committee capable of addressing ICT concerns Funding for engagement activities Tools to support the ICT community Database of project 'lessons-learned'	**What needs to be done to ensure social desirability?** Create incentives for thinking about research outputs Encourage appropriate development approaches **What training is required?** Understanding of regulation **What infrastructure is required?** Accessible participatory tools and methods Open access to data and publications	**How do we ensure that the implied future is desirable?** Consider implied future state at project/programme inception **What training is required?** Understanding of ELSI Understanding current debates and controversies **What infrastructure is required?** Reflection on purpose part of funding mechanisms Reflection on purpose part of project evaluation criteria	**Who matters?** Stakeholder participation **What training is required?** Contextualise projects as sociotechnical **What infrastructure is required?** Community building Leadership council Champions / Advocates Science education to allow the public to engage intelligently

(Source: Orbit, 2020)

- Governance (robust, adaptable, and permitting structural changes).

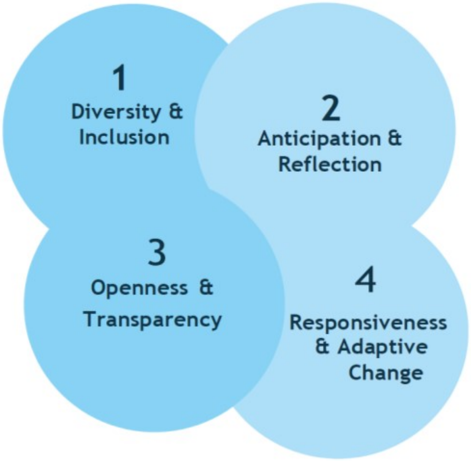
- Open access (information resulting from publicly funded research).

- Public engagement (more open and inclusive R&I).

- Science education (education focused on societal needs and participation of citizens).

c. Adopting specific process dimensions (Figure 7-2):

Figure 7-2 Process dimensions



(Source: RRI Tools, 2020)

- Engagement, diversity, and inclusion of various stakeholders (from the beginning).

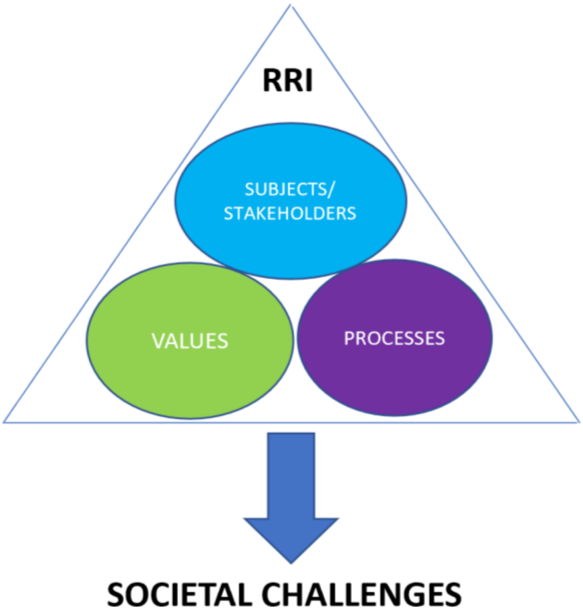
- Anticipation of needs of various subjects, reflection on future effects, sustainability, motivation, purposes, and uncertainties. This includes considerations on methodology, risk assessments, and ethical approval.

- Openness and transparency, which increase accountability and research integrity.

- Action, responsiveness, and adaptive change (reactive and flexible to new knowledge and circumstances).

Therefore, RRI is a multi-dimensional framework inclusive of different components (Figure 7-3). To deal effectively with societal problems, and to anticipate how research can influence the future, RRI calls for the integration of social values, the adoption of specific processes, and the involvement of different subjects and stakeholders.

Figure 7-3 RRI as a multi-dimensional framework



(Source: Chiara Addis)

In the following section, the specific RRI tools will be presented.

7.2.1 RRI tools

The *RRI Tools* is a European multidisciplinary consortium created to exchange RRI best practice amongst organisations and people. Its members are "active in this new vision of scientific and social development" (RRI Tools, 2020) and can use the toolkit and contribute to it. Some of the suggested practices are of interest for this research, as they identify and promote a modus operandi between different factors. For example, the link between inclusion and creativity, with the importance of involving various stakeholders facilitating the evaluation of ideas from different perspectives. Another suggestion regards the assessment of innovation carried out considering various elements, for example:

- Positive effects (usually more easily identified), negative effects (usually less predictable), multiple effects in various areas (e.g., socioeconomic impacts), and time of the assessments (pre and post innovation).
- Direct causalities and other indirect elements, with impact paths and connections amongst different factors, e.g., the chain of effects (inclusive of outside factors e.g., market dynamics).

These examples of best practices can prove useful in the management of AI and DP, and they will be better discussed in the model presented in Chapter 8.

7.2.2 RRI, privacy and DP

The amount of research on the relationship between RRI, privacy, and DP has been moderately small. However, two works by Stahl (B. C. Stahl, 2012; B. C. Stahl & Wright, 2018) are especially relevant for this research. The first paper (B. C. Stahl, 2012) is an insightful and forward-looking analysis on RRI and privacy/DP conducted long before the GDPR. He considers privacy as a key area within RRI. Noticeable are some considerations made on the privacy implications of future technologies and "on the changing nature of privacy preferences" (ibidem, p. 721). Listing a number of possible tools and activities to assess privacy risk, he identifies some activities that are now GDPR requirements, such as the DPIA and privacy by design. He also envisages an RRI implementation inclusive of regulations, institutions, and specific guidance on technologies to "develop principles and standards of good practice and integrate these into research and innovation processes" (ibidem, p. 713). Relevant is also his understanding of privacy as related to different aspects of RRI (activities, actors, and normative foundation), and his view of RRI as a meta-responsibility, related to existing responsibilities and not as a new type of responsibility.

His analysis of the relation between privacy and future technology, his understanding of privacy as a concern of different areas, and the need to see privacy not as a new responsibility, all show a significant understanding of privacy as a phenomenon that is a pervasive feature within organisational processes in need of a holistic approach.

The second paper (B. C. Stahl & Wright, 2018) focused on RRI, AI, Ethics and privacy. After presenting different existing initiatives and regulations that can enhance ethics, the authors introduce RRI as a framework able to foster collaboration and create synergies. The case of the EU's Human Brain Project (HBP) is provided as an example of the integration of RRI into research (e.g., ethics management), and of what could be improved (e.g., engagement with communities). While supporting the use of RRI in research, the authors are also aware of some of its limits, such as its adoption mainly being limited to universities and publicly funded research.

7.2.3 RRI, management and industry

The relationship between RRI (originating in the public sector), management, and the industry is complicated. The engagement of innovation management and governance researchers and practitioners with RRI is limited (Owen, 2014). RRI is considered “first and foremost a process innovation, which can offer 'a new mental model for innovation policy and its delivery, a potential paradigm shift’ ” (ibidem, p. 3).

Drawing attention to Horizon 2020, and its aspiring visionary science and innovation, Owen observed some difficulties. The difference between *social desirability* and *social acceptability,* between what is hoped for and what is possible, is often significant. Furthermore, innovation often comes with a certain amount of uncertainty and ambiguity, often intensified by organisational practices, such as the “problem of many hands” (ibidem, p. 4), which occurs when most activities are split between multiple individuals, impacting on responsibilities and accountability. Another key point is the control of innovation, for example via regulations, which is often minimal at the early phases. Whilst regulating is important, regulating prior to knowing the impacts of the innovation can risk a “lock-in” of the innovation (ibidem, p. 5). Other useful observations are related to the need for more context-specific elements, and the possibility that RRI would represent an extra cost for small companies that are focusing on remaining commercially viable.

The elements highlighted by Owen recall some of the findings from the case studies. For example, the debates around the pace and regulation of AI (Experts), the silo mentality approach in the management of data within organisations (CS1), and the perception of the GDPR as top-down regulation impacting on innovation but not on culture change (CS2).

Two other important papers for this research focused on RRI and Business. The first is a paper published by the University of Salford which looked at how RRI could be combined with the Knowledge Transfer Partnerships (KTPs), a UK scheme that facilitates innovation and productivity in UK businesses (Kreps et al., 2016). The paper is an interesting example of how RRI could be applied with other pre-existing frameworks in order to involve businesses. The second is a paper published by Stahl (2018) that looked at the specific challenges faced by businesses in their use of RRI. Observing how private companies create most of the innovation (especially in the ICT sector), he raised the question of the suitability of a framework created within the public research sector, which does not include the innovation resulting from creative destruction, nor provide the more tangible benefits required by the industry. And yet, he observes how some elements of the framework were deployed successfully in projects within the private research sector, for instance, with regard to the engagement of users and stakeholders, and the development of more interdisciplinary and integrated project management. Both factors are presented as key elements for the realisation of successful projects integrating RRI elements.

7.2.4 RRI in practice and recent development in AI ethics

Particularly interesting for this research are also two reports which focus on the application of RRI, and a paper on the most recent developments around AI ethics.

The first report was published by the RRI Tools Consortium which analysed the opportunities, obstacles and needs that emerged in RRI practices in Europe (Smallman et al., 2015). The report highlighted some important findings regarding the role of stakeholders and industry, some of which recall those identified in the case studies presented in this research. Stakeholder dynamics were reported as key elements in RRI practices, and some common challenges were identified; for example, frequent lack of collaboration, actions not stakeholder specific, responsibilities often assumed to be falling on other stakeholders. Business and industry were also described as having some specific characteristics. Being focused on profits, businesses need to see the advantages and the ROI before accepting the perceived associated costs of applying RRI. Furthermore, the RRI framework is often perceived as abstract, superficial, and lacking flexibility. While RRI aims are clear, the indications on how to achieve them or who is in charge are often less clear. Another important point referred to the perceptions of individuals, with top-down and paternalistic tones capable of impacting the buy-in of the framework. Noticeable is how some of the findings of the report mirror some considerations expressed by the research participants implementing AI and DP in this study (CS2), and also some experiences of organisations operating in very fast-moving markets (Experts).

The second report was published by SHERPA Project (Shaping the Ethical Dimensions of Information Technologies – a European Perspective) (Brey et al., 2019), an EU Horizon 2020 project on the ethical use of AI and Big Data. The report is the most important and comprehensive document on the implementation and use of AI, a rare work focusing on the operational aspects of the technology and on its ethical use. The document was extremely important for this research, and it clarified some key aspects, such as the multiplicity of operational ethics requirements, and how ethics can be applied to the full life cycle of innovation. And in the light of this research, various elements in the document were critically analysed. This report provides guidelines specifically aimed at those roles having ethical responsibilities (e.g., ethics officers) and it dedicates to privacy (which implicitly includes DP) a small part. It mentions the GDPR, issues regarding data quality, integrity, and access. It does not make any reference to the Data Protection Act 2018. While performing ethical assessments and creating diversity reports is recommended, performing DPIAs is not mentioned. Human oversight is considered crucial for exercising control in augmented systems, and overconfidence and over-reliance on the system are clearly considered significant risks in such models. The document links diversity and fairness mainly to algorithm biases and recommends using AI systems "with an inclusionary, fair, and non-discriminatory agenda" (ibidem, p. 6). It recommends stakeholder engagement and the involvement of developers from diverse backgrounds. However, similar provisions are not made for those using the system, nor for other roles within the organisation. Similarly, the report identifies a risk in external companies both processing and storing personal data and indicate this as a factor that should concern developers. Other roles within the organisations are not mentioned. Describing the models for the ethical use of AI, the report makes a distinction between IT governance and IT management. IT governance is considered the responsibility of the Board, that provides the strategic direction, while IT management refers to the management of the operational IT, responsibility of the executive management. Other roles within the organisation, other areas, or other internal stakeholders are not discussed as having some responsibility around AI ethics. For example, DPOs are briefly mentioned when recommending their involvement in the development process. No involvement of the DPO is envisaged in the implementation or use of AI, nor BAU. Similarly, other internal roles or subjects who have or might have an interest in AI or privacy are not considered. Responsible AI seems to be mainly confined to the role of the Ethics Leader and to those in charge of IT, who "adopt and implement relevant ethical guidelines for the IT field" (ibidem, p. 9).

Therefore, AI appears to be conflated with IT, and IT management is considered sufficient to manage AI/ML characteristics. For example, the report recommends testing the business case on ethics guidelines. This is clearly valuable. However, this implicitly considers AI projects as having a clear lifecycle, with a defined business case at the beginning. But this does not always appear to be the case, as seen by the growing trend of repurposing, with AI systems being used for other purposes once inside the organisations. Furthermore, this internal top-down approach focusing on IT seems to mirror the external one. The principles of the framework originated at the EU level, external to specific industries, and similarly to the GDPR they could be perceived by organisations as being imposed by distant institutions.

Organisations do not appear to have an active role in co-creating any ethical content. IT management is called to "encourage a common culture of responsibility, integrating both bottom-up and top-down approaches to ethical adherence" (ibidem, p. 13). And yet, the perceptions, experiences, and knowledge of individuals, and their role as active agents able to share and create ethical values, appeared to be overlooked.

Therefore, some aspects of the implementation and use of AI appeared to be disregarded.

A more recent paper (Ryan & Stahl, 2020) is specifically focused on AI, ethics, and the use of technology. The authors analysed a high number of reports and papers, aiming at translating existing AI ethical guidance into practical suggestions specifically for developers and users. They highlighted some rules in key areas and indicate (in a quite normative and direct way) what should be done in those areas. For example,

- Transparency. Organisations should perform various checks (e.g., algorithm and data auditing, and monitoring of outputs).
- Justice and fairness (e.g., control of the level of justice in the system, use of unbiased or incorrect data, inclusion, non-discrimination, diversity, accessibility to data, remedy, and redress).
- Non-maleficence (e.g., security, safety, prevention).
- Responsibility (inclusive of the "responsibility gap" in augmented systems, accountability, and liability).
- Privacy and personal data.
- Beneficence (e.g., social good, peace).
- Freedom and autonomy (e.g., consent, self-determination, empowerment).
- Trust, sustainability, dignity.

The paper is extremely comprehensive and valuable. It provides precious help for those who are navigating the vast amount of published resources on AI ethics and are looking for practical guidelines more focused on the use of the technology. And yet, as addressed by the authors at the end of the paper, something is missing in the current discourse on ethics. All the rules included in those documents indicate what should be done, but not how.

Both the report published by Sherpa and the paper by Ryan and Stahl shifted the focus from the development of AI to its use, and this change was very much needed. But the crucial question is how to translate those societal and ethical values into organisational practice. For example, when people have to choose between different trade-offs, how do they address the conflicts? This is the moment of choice when individuals face a dilemma and have to mediate and choose between different options. Those moments can lead to different outcomes. For instance, they can translate into a cost (which organisations generally want to avoid), have a huge influence on strategic decisions and risk management, and can produce unethical consequences. How do people choose in those specific contexts? In those specific creative milieux? How is their use of AI and personal data informed by their knowledge, experience, and power relations?

This research attempts to provide some answers. However, in order to do so, other conceptual elements are needed to complete this theoretical framework, such as power and the experience of people. Those are provided by the work done by Andrew Feenberg and critical theory of technology (CTT). His work was important in the development of RRI (Hassan, Mingers, & Stahl; 2018), and some elements in his theory can provide further insights for this research.

In the following section, different approaches within critical theory will be presented, firstly with the debate around critical research, and secondly, with the presentation of CTT as an alternative paradigm for this research. The work of Feenberg will be then presented in detail.

7.3 Different approaches within critical theory

There has been a long debate about the notion of criticality within IS giving rise to different concepts and fields of study. In order to provide some clarity and explain the position adopted by the researcher, a brief section on the main debates occurring around critical discourses in IS is presented in this part, explaining some key concepts within critical theory, focusing upon critical research and critical theory of technology (CTT).

7.3.1 Critical research in IS

Critical theory in IS is an umbrella concept used for various theories that have critical approaches and methodologies (Zheng & Stahl, 2011). Within critical theory, Zheng and Stahl identify two critical discourses, critical research, mainly used by IS scholars, and CTT, more common among philosophers of technology. Critical research in IS is concerned with "identifying power relations, conflicts and contradictions, and empowering people to eliminate them as sources of alienation and domination" (Oates, 2005, p. 296). Critical research in IS has been characterised by an ongoing debate about the question of what is the most appropriate research perspective for studying the phenomena related to IS. The debate has been long and complex and some of the key positions are summarised below.

A now famous paper by Orlikowski and Baroudi (1991) started questioning the common theoretical perspective in IS. After inquiring about the dominance of positivism, the authors classified the theoretical perspectives according to three different orientations: positivist, interpretive, and critical. While positivists and interpretivists explain reality, critical researchers aim to transform it by criticising and identify contradictions in its structures. Social reality is seen as a human product, but also possessing some objective elements. As economic, political, and cultural structures dominate and alienate humans, critical researchers aim to raise awareness, support the elimination of such domination, and create different social orders. Class and socio-economic factors are considered key factors in shaping social relations, with the antagonist relation between management and labour being a distinctive factor of the capitalist mode of production. The authors also identified some important limitations in critical research. For example, the lack of self-reflexivity among critical researchers around their own concepts and theoretical models, or the prevalence of determinist views that assume management to be craving more control and employees to be completely ignorant or lacking means to eliminate their oppression.

Chen and Hirschheim (2004) carried out a similar overview of existing research and observed that IS research was still based on a positivist approach, even if more qualitative work had been produced since the paper by Orlikowski and Baroudi. Their survey did not include any critical papers. Such lack of critical works was strongly criticised by Richardson and Robinson (2007), who identified a small but growing field of critical research in IS. Critical research was seen as difficult to define. While traditional functionalist approaches aim at understanding IS, critical research aims at criticising the status quo, identifying structural contradictions, transforming restrictive and alienating social conditions, and promoting emancipatory social change. However, the difficulties in clearly defining emancipation and critical information system research (CISR) were also noted by the authors. Kvasny & Richardson (2006) stressed the political agenda of critical research, its empirical sensitivity, the uncovering of institutional repression and resistance, and the combination of theory and praxis. Emancipation as the realisation of needs and potential is also central in Myers and Klein (2011). Zheng & Stahl (2011) stressed how critical intention (the perception that social reality can be changed and improved) and emancipation are key concerns in critical research. They highlighted the influence of Marx, the Frankfurt School, and Foucault. They further showed how CTT, too, was influenced by Marx and the Frankfurt School. However, as we will see in Feenberg's work, Marcuse and Heidegger are also extremely influential. Considering CTT's conceptual stances (i.e., the nature of technology as something that is shaped and constructed), this is more closely related to the philosophy of technology. Critical research in IS was also influenced by other critical philosophers, such as Bourdieu, Foucault, and Latour. (Oates, 2005).

The next section discusses the interconnections and some of the differences between critical research and CTT in order to explain the researcher's preference for CTT.

7.3.2 Critical theory of technology as an alternative paradigm

While critical research provides a highly politicised framework for empirical research around a tight and rigid vision of the role played by people within organisational contexts, CTT assumes a more philosophical approach which has a higher degree of flexibility, allowing for a stronger consideration of human agency. This renders CTT a more suitable paradigm for framing this research than critical research, in particular if combined with RRI. This and the following sections are dedicated to explaining why CTT was preferred over critical research as an important element in the theoretical framework.

As illustrated in closer detail in the section on Feenberg's work, subjects in CTT are not only oppressed. They are also described as active agents shaping systems and technology. In particular, CTT's critique of positions that are quasi anti-modern in both the philosophy of technology and social studies of technology felt very appropriate to a researcher that strongly believes that AI can, and should, be used responsibly. The idea that technology is ambivalent but not neutral, and that it can be separated from an original context and then adapted to a new one, perfectly fitted the issues surrounding the development and use of AI explored in this research. CTT comprises the view that technology is socially relative and that technical rationality is not context-free. Decisions are then more dependent on hegemonic values (and less on an ideological expression of class interests). This perspective was considered to be more suitable for explaining how AI was used in different contexts. Furthermore, the concepts of power, strategy and tactics provided more refined tools for analysis than the concepts of oppression and domination.

Even though some of CTT's explanatory elements (i.e., power and the management-staff relation) may be in need of further elaboration in order to account for highly differentiated ecosystems where different stakeholders connect via complex relationships, CTT was considered a richer and more suitable theoretical choice than critical research.

In the next section, some of the key insights of CTT useful for this research will be presented.

7.4 Feenberg and critical theory of technology

Critical theory of technology (CTT) was created by Feenberg drawing upon the philosophy of technology and constructivist technology studies (Feenberg, 2005). Influenced by the Frankfurt School, Heidegger, and social constructivism, his theory considers technologies and technological systems at different levels, and it is "both critical and empirically oriented" (ibidem, p. 62). He considers technology as "socially shaped and constructed" (Stahl & Wright, 2018, p. 73) and instrumental for modern hegemonies (Feenberg, 2005).

CTT aims at understanding technology through the analysis of the context, its conditions, its values, and power relations. Technology is viewed more as a process than "a thing", not neutral but ambivalent, suspended between different possibilities. "This ambivalence is distinguished from neutrality by the role it attributes to social values in the design, and not merely the use, of technical systems." (Feenberg, 2002, p.15)

Through critical theory, Feenberg identifies how invisible codes condition "values and interests in rules and procedures, devices and artefacts that routinise the pursuit of power and advantage by a dominant hegemony" (ibidem). When effective, hegemony is not imposed, as it is "reproduced unreflectively by the standard beliefs and practices of the society it dominates" (ibidem, p. 75). CTT principal characteristics are presented below.

7.4.1 Technology and the subject

The role of the subject in relation to technology is an important element that differentiates Feenberg from Marx, Heidegger, the Frankfurt School and Marcuse. The "impersonal domination" (Feenberg, 2005, p. 47) of the subject, common in the philosophy of technology of Marx, Heidegger and the Frankfurt School, is considered by Feenberg too abstract and inadequate to understand contemporary phenomena. From the hostility to technology of Adorno and Horkheimer to the emphasis on the potentialities in technological development showed by Benjamin and Marcuse (Feenberg, 2002), all these explanations remain unsatisfactory for Feenberg. Marcuse considered technology as embedding class divisions, as a means for domination with "a one-way direction of cause and effect" (Feenberg, 2005, p. 48). Even though he was later more open to the democratic potential of modern technology, his approach is still unsatisfactory: "none of these positive evaluations of technology are sufficiently developed to intersect fruitfully with contemporary technology studies" (Feenberg, 2002, p. 34)

Observing the role subjects play within the systems, their interactions, and the feedback they receive, Feenberg can see an active exchange between the subjects and the systems, which he defines as "a practical significance of embodiment" (Feenberg, 2005, p. 48).

Noticing how those subjects excluded from the design of the technology are generally those who suffer the most from its consequences, he envisages their involvement in the design as a "democratic transformation from below" (ibidem, p. 49). This position recalls the need to involve subjects/end-users in the development of AI as a measure to increase fairness (as seen in 2.4.2). Could the involvement of the subjects in the implementation and use of AI produce a similar transformation from below?

7.4.2 Instrumentalization theory

Feenberg's idea of technology is both critical and appreciative of the philosophy of technology and social study of technology. The philosophy of technology is in general viewed as "abstract and unhistorical" (ibidem, p. 49), mainly essentialist and quasi anti-modern. Conversely, while social studies of technology are appreciated for their rich complexity, they are thought to ignore "the larger issue of modernity and thus appears uncritical, even conformist, to social critics" (ibidem, p. 49). And yet, he does not see them completely in opposition, and he draws upon both to elaborate his critical theory of technology. He distinguishes between technologies embedding technical rationality from technologies embedding underlying societal values, interests, and priorities (Hassan et al., 2018). In order to understand technology, Feenberg creates his Instrumentalization Theory which considers two different levels:

1. Primary instrumentalization: the technological object is decontextualised/separated from the context and considered for its primary and useful properties. Although being inspired by categories introduced by Heidegger and other substantivist critics of technology (ibidem), they are not considered by Feenberg from an essentialist point of view. This is the "systematic reductionism…of technical choices, codes, and designs, before these encounter the further social worlds of users and reactions…" (Feenberg, 2010, XIV).

2. Secondary instrumentalization: the object is re-contextualised/adapted to a specific context. New elements (e.g., values) from that new context can be integrated into the design. "This involves a process which, following Heidegger, we can call 'disclosure' or 'revealing' of a world" (ibidem, p. 50). This level of analysis is inspired "by [the] empirical study of technology in the constructivist vein" (ibidem, p. 51). Feenberg argues that it is crucial how the subjects understand the devices and the systems, and how they perceive them.

Therefore, he does not reject categories of traditional social theory, nor completely accept an "exaggerated and largely rhetorical empiricism… to integrate its methodological insights to a more broadly conceived theory of modernity" (ibidem, p. 51).

7.4.3 Culture and technical codes

Feenberg argues that criteria used to evaluate the success of a system are strictly dependent on culture. More determinist and instrumentalist approaches consider efficiency as the key criteria for evaluating the success of a technology. Philosophy of technology tends to reject the universality of rational conditions. For example, constructivists show how the success of a system can be attributed to different options:

> [t]he different interests of the various actors involved in design are reflected in subtle differences in function and preferences for one or another design of what is nominally the same device. Social choices intervene in the selection of the problem definition as well as its solution (Feenberg, 2005, p. 51)

Therefore, the rational efficiency of a system is no longer sufficient to define its success. If that system also generates negative consequences, it cannot be considered a success. Feenberg sees a strict connection between the subjects involved in the design of the system, and the choices of needs, solutions, and characteristics of the resulting design. Their interests and characteristics are reflected in that design. Thus, technology cannot be considered rational. Depending on a choice between various interests and ideologies, technology is "socially relative" (ibidem).

> If there are no context-free universals, context-freedom cannot serve as a meaningful standard of validity and so the failure to meet that standard does not imply relativism but rather the pragmatic principle of openness to revisions and dialogue with those whose contexts differ (Feenberg, 2017)

Feenberg elaborates the concept of Technical Code to explain how different options are chosen: "Critical theory shows how these codes invisibly sediment values and interests in rules and procedures, devices and artifacts that routinise the pursuit of power and advantage by a dominant hegemony" (Feenberg, 2002, p. 15). Thus, the choices (and decisions), taken in context are more dependent on the hegemonic value system operating in that environment than on universal criteria. Therefore, the technical rationality leading to a choice is:

> neither an ideology (a discursive expression of class interest) nor is it a neutral reflection of natural laws. Rather, it stands at the intersection between ideology and technique where the two come together to control human beings and resources in conformity with what I call 'technical codes' (ibidem, p. 15)

7.4.4 Operational autonomy and technocracy

Feenberg refers to Operational Autonomy as the power of the roles in charge to decide independently from other subjects (e.g., employees and other external actors). On the one side, this freedom generated a variety of new values and requests. On the other side, "ethical demands forced to seek voice discursively and realisation in the new technical arrangements" (Feenberg, 2002, p. 22). At the same time, technocracy concerns the expansion of power based on technology and management to societies. This perpetuates the rational domination by a small group. The assumed rationality and neutrality of technology are once again refuted by Feenberg.

7.4.5 Technology, power, and resistance

Technology is the result of technical knowledge and power structures, and different structures create different innovations and produce diverse social consequences.

Feenberg envisages democratisation of technology, which comprises the involvement of new and disempowered subjects. But the mere involvement is not sufficient, as their power and agency need to be supported in order to resist the dominant management. He makes a distinction between those who rule and those who are ruled, clarifying how “[s]ubordinate actors must intervene in a different way from dominant ones” (Feenberg, 2005, p. 55).

In defining his vision of resistance, he is strongly inspired by Ihde (1990) and de Certeau (1984). Ihde and Feenberg are both influenced by phenomenology and, as Feenberg pointed out, they belong to the “empirical turn” in the philosophy of technology (Feenberg, 2015, p. 1). Ihde argues that “the crisis of modernity must be overcome through a ‘gestalt switch in sensibilities [that] will have to occur from within technological cultures’” (Feenberg, 2017, p. 115). This switch is possible because “technologies do not stand alone. They are always interpreted and employed in a cultural context. The ‘multistability’ of technology holds open the possibility of change ‘from within’” (Feenberg, 2015, p. 2). Agreeing with Ihde, Feenberg believes that change does not result from inputs initiated externally (e.g., within politics, philosophy, religion) but needs to be created from inside.

Conversely, de Certeau focuses on everyday practices that ordinary people adopt in order to subvert rules and obligations imposed upon them, people that “may be able to discover creative activity where it has been denied that any exists” (De Certeau, 1984, p. 167). His understanding of power, inspired by Foucault, offers Feenberg the distinction between strategy and tactics. Strategy belongs to groups exercising their power from their institutional base (e.g., managers and state administrators), while tactics belong to other groups who, lacking a base for acting continuously and legitimately, need to use their power via micropolitical resistance (Feenberg, 2002). That could challenge the technically based hierarchy, as “[s]ince the locus of technical control influences technological development, new forms of control from below could set development on an original path” (ibidem, p. 16).

Feenberg advocates for a democratisation of technology that privileges the “excluded values and the publics that articulate them” (ibidem, p. 22). This can happen via a decontextualisation-recontextualisation of the technology that permits the inclusion of these values into the design.

7.4.6 Recontextualizing strategies

Technology encompasses both the idea of efficiency and the ideals operating in the past contexts where it was created. “[T]he division between what appears as a condition of technical efficiency and what appears as a value external to the technical process is itself a function of social and political decisions biased by unequal power” (Feenberg, 2005, p. 57).

Thus, technologies embed past decisions, which were the results of past trade-offs between different values/options and different actors. For this reason, for Feenberg the trade-off between efficiency and values is a false dichotomy. New functions, inclusive of new subjects’ requests can be taken into consideration and added to existing functions, without impacting on the technical efficiency.

7.4.7 Terminal subjects

As seen, although being hugely influenced by Heidegger and Marcuse, Feenberg considered their positions inadequate for understanding new phenomena. For instance, the diffusion of the computer, and how people have been appropriating the medium, have increased users ’ communication and helped them to create online communities. Computers could not only be considered negatively, for example, as capable of dominating humans or a cause of human communication degradation (as per the “post humanist” approach inspired by cultural studies). Those approaches neglect the active role of the user: “Approaches based on modernity theory are uniformly negative and fail to explain the experience of participants in computer communication. But this experience can be analysed in terms of instrumentalization theory” (Feenberg, 2005, p. 59).

Feenberg did not see humans as mere “terminal subjects”, lacking agency and being easily controlled. They are active agents who use technology as a tool and create new communication practices, and their experiences have a central role. In failing to see this active role, previous approaches also failed to see how technologies can be transformed. This is the central question in his theory: how actors are influenced? How do they interpret and shape the design and use of the technology?

As seen, operational autonomy is instrumental to those who rule. But those who are ruled can acquire a new autonomy “that works with the ‘play’ in the system to redefine and modify its forms, rhythms, and purposes…a reactive autonomy [called] ‘margin of maneuver’” (Feenberg, 2002, p. 84).

How the technology is used then is not just a question of mere rationality. What is done by the users, their experience, and how they adapt to the technology all have an impact on designers, their actions and intentions (Kreps, 2019). Both the designer of the technology and its user are relevant for Feenberg, as both are significant forces in shaping the technology.

Therefore, Feenberg views technology, subject, values, and context as deeply connected.

The subject is not passive but active in interacting and shaping the technology in a specific context, an environment where systems of values operate, and where oppression, resistance and negotiations between different values can impact on strategies and technology.

This approach is particularly relevant for this research, where AI is analysed when implemented in specific contexts.

In the following section, the elements from RRI and CTT used for the new critical framework will be presented in closer detail. The section then ends by showing how data can be read through it.

7.5 The new management framework: critical AI&DP/*CRAIDA*

In order to explore the praxis of AI, DP and FAT principles, this research draws upon some specific elements of RRI and critical theory of technology which can help to better read the relationships between subjects, technology, and organisations.

As seen, RRI provided a suitable framework for improving R&I by satisfying societal challenges and considering their ethical dimension. It further contains elements for empowering different actors and provides a more holistic and collective approach but can be abstract and lack flexibility. Feenberg, on the other hand, provided suitable tools for exploring the role and the experience of subjects inside the organisations. He sees the subjects (designers and users) as capable of influencing the technological objects in the design phase of the innovation, while this research looks at how AI is influenced and shaped by the subjects during its implementation, and how processes, context, dynamics of power, values, understanding and perceptions all impact an ambivalent technology.

Considering the general discourse on the development of AI, the novelty of the GDPR, and the focus on data and FAT, this research project required a different approach to consider other aspects being left out, such as the implementation of AI.

Thus, drawing upon RRI and CTT, the researcher defined a critical AI and DP framework that could take full account of the agency of humans (as users and developers) in implementing ethical AI and DP practices.

The approach is informed by the data collection analysis carried out as part of the PhD. The framework aims at identifying the elements shaping the implementation of AI, the values influencing AI, how the decisions are made, and under which conditions and assumptions. Specifically, *CRAIDA* includes the following elements:

a. **Societal Values** (RRI). The societal values considered in the organisation, their understanding, and their perception (i.e., fairness, diversity, inclusion, openness, transparency).

b. **Context and Culture** (Feenberg). The influence of the context onto the praxis.

   For example, external factors (e.g., regulations, sector, industry, market) and internal factors (e.g., organisation characteristics, rules, culture — with hegemony and technical code).

c. **Subjects' experience** (F). The influence of the subjects on the praxis (their knowledge, understanding, perception, and performed role).

d. **Power** (F). Power and power dynamics, inclusive of strategies of leaders and managers, with micropolitical forms of resistance via informal tactics.

e. **Risks** (RRI). Identified risks, how they were assessed (e.g., future positive/negative consequences), and specific AI/ML risks.

f. **Processes** (RRI). Processes and innovation. e.g., management of information (practices and challenges, e.g., silo mentality, the problem of many hands); implementation of new regulations and services; attitude to new events (new inputs for innovation) and process of re-contextualisation with that inclusion of new elements/values.

g. **Stakeholders** (RRI). External and internal stakeholders (e.g., individuals, teams, entities), reasons for the inclusion/exclusion and effect of the collaboration.

h. **Decision-making** (F). Decision-making process, criteria, and tools used to make choices, from the choices around the system to the choices made by the system (e.g., efficiency/other values, cost efficiency, mediation/dialogue).

Other elements of RRI and CTT were not considered useful for this approach. For example, ethics in RRI refers to research integrity and ethical acceptability of the outcomes. At the same time, the current research on AI ethics is mainly focused on the integrity of data and algorithms. By considering the first as an inspiration, and the latter as a limited interpretation, this research expands AI ethics into the more comprehensive and holistic interpretation of responsible AI management.

Some of the RRI key societal values were not included in the critical AI and DP approach. For example, gender equality was not considered relevant in specific cases per se. It was, however, considered within the more inclusive concept of diversity as an over-arching concept taking account of multiple differences (as also seen in CS1). Other values, such as open access or public engagement were considered to be less relevant for this research.

Similarly, the concepts of Operational Autonomy and Technocracy in Feenberg were not considered to be useful for reading the reality of two organisations using AI to empower and support others, as they imply the rational domination of a small group on others.

Table 7-1 includes the elements from RRI and Feenberg included in and excluded from the *CRAIDA* framework. Therefore, the *CRAIDA* theoretical approach aims at identifying the elements shaping the implementation of AI and DP, the values influencing AI, how the decisions are made, and under which conditions and assumptions. Unveiling how these elements are interconnected can help to identify strategies to reduce the potential risks posed by AI systems processing personal data.

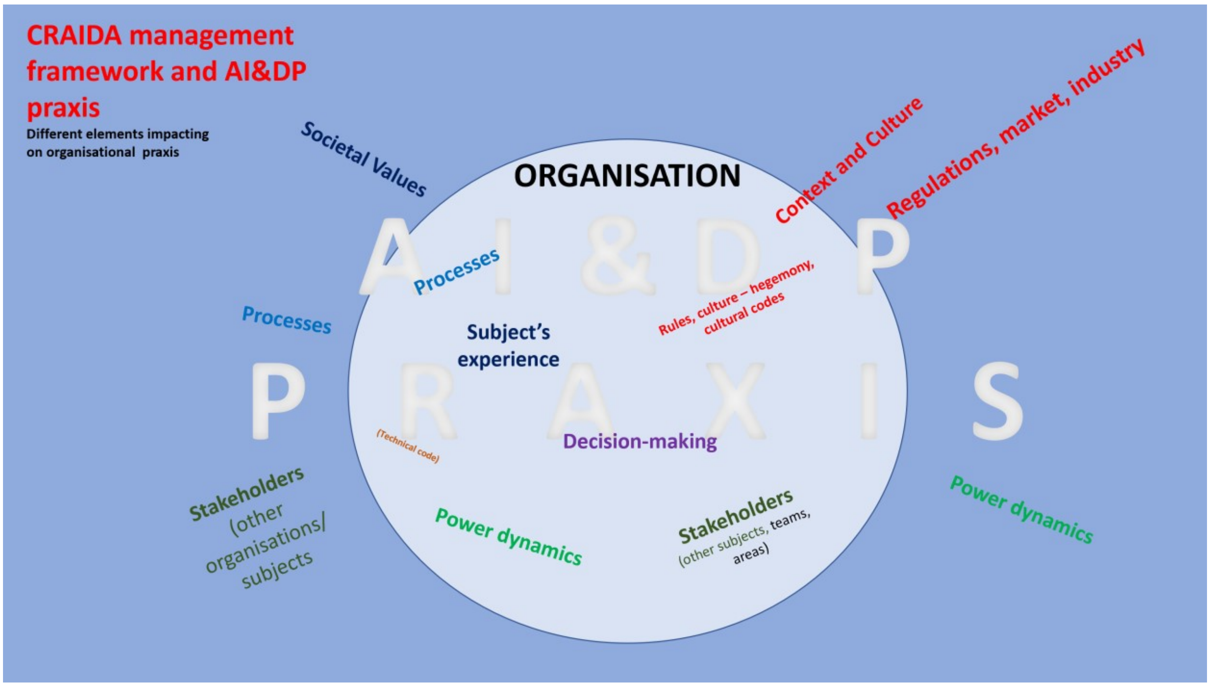
Table 7-1 RRI and CTT elements included in CRAIDA

Included	**Not included**
RRI	**RRI**
Societal Values (e.g., diversity)	Specific vision of ethics (e.g., research integrity). *CRAIDA* has a more holistic and bottom-up approach to values

	Fairness linked to algorithm biases (this research expands this, including other factors, as per Chapter 8)
Risks (identification and assessment)	Specific societal values (e.g., gender). *CRAIDA* is more intersectional
Processes (link to innovation, practices, and information management)	Structured processes and stakeholder dynamics (low flexibility, more for public institutions)
Stakeholders (external, internal, reasons for including/excluding them)	Open access (not relevant for this research)
AI ethics as a strategic element	AI ethics linked to specific roles, and IT governance and operational IT (this research expand AI ethics/responsibility to other roles)
	Public engagement is not relevant for this research (but the end-user involvement is relevant)
Feenberg/CTT	**Feenberg/CTT**
Subject experience (influence of subjects on praxis)	Operational autonomy and technocracy, domination of few (as it refers to management that dominates others)
Context and culture (influence of context and its values, technology and praxis)	Power of few and resistance (bottom-up). A dualistic vision of rulers and ruled
Power (dynamics, agency in different roles, micropolitical forms of resistance via informal tactics)	
Decision-making (process, criteria, tools)	
Recontextualizing strategies. Technology encompasses efficiency and values for past context, then new adaption/recontextualization (important for Augmented AI)	

Figure 7-4 shows the various elements of the *CRAIDA* theoretical approach in relation to AI and DP praxis.

Figure 7-4 CRAIDA management framework and AI and DP praxis


CRAIDA management framework and AI&DP praxis
Different elements impacting on organisational praxis
Societal Values
ORGANISATION
Context and Culture
Regulations, market, industry
AI & DP
PRAXIS
Processes
Processes
Subject's experience
Rules, culture – hegemony, cultural codes
(Technical code)
Decision-making
Stakeholders (other organisations/ subjects
Power dynamics
Stakeholders (other subjects, teams, areas)
Power dynamics

(Design: Chiara Addis)

The researcher sees this research and the *CRAIDA* management framework as part of the critical theory tradition (Stahl, 2008). The research has the critical intention of initiating and promoting change within organisations. This is done to increase responsible management practices, promoting the ethical use of AI, and encouraging the active involvement of different subjects in the innovation process. It draws upon CTT, which originated within the critical theory tradition (i.e., in the tradition of the Frankfurt School) and it uses critical methodologies that are typically utilised by critical researchers (qualitative and reflexive) (ibidem, p. 11).

7.6 Using the *CRAIDA* management framework

The *CRAIDA* management framework was used to critically read the empirical data and identify the elements shaping the implementation of AI, the values influencing AI, how the decisions are made, and under which conditions and assumptions. The section below presents the results, considering the elements of the framework in 7.4.

a,b,c Societal values, context and culture, and the experience of the subjects

As seen in the interviews, the awareness around unethical uses of AI is growing inside organisations thanks to the debates in the public domain, and the work carried out by those who are more aware and knowledgeable, such as progressive leaders and vendors investing in research. As the debate grows, it is critical to see how organisations were dealing with the topics and how participants viewed and perceived some of those issues. The internal debate revolves around some key areas. For example, one area concerned the strategic choices made by organisations. The desire and commitment of some organisations to invest in responsible AI are growing. This increasing interest in responsible AI is less visible in the general discourse on AI, which is generally focused on the exploitation of personal data. However, the desire to use AI to do good is very strong in both case studies and some of the interviews with the experts. Fairness is central in both case studies, and compliance is considered instrumental in doing good. Both organisations are developing innovative projects and creating ethical systems. Similarly, some of the organisations mentioned by the experts had altered or stopped AI projects for ethical reasons. The desire to do good also appeared to be a long-term plan in both projects. And yet, the investments made to support the sustainability of ethical practices appeared to be less clear. Additionally, having an ethical aim does not eliminate the risk of producing unfair outcomes during the implementation or the use of the system, nor does it exclude a later repurposing of the system for more unethical aims once in use.

Another area was related to the FAT principles and the different meanings ascribed to the concepts. The principles are not only key GDPR requirements for processing personal data, but they are also societal values not limited to the realm of DP. When using AI “for good”, the understanding of fairness, accountability, and transparency during the implementation and use is particularly important. How people inside organisations understand them, in those specific contexts, and how those values are being translated into their management practices and envisaged in future uses of AI, merits much more attention.

The concept of fairness was not always ascribed a unique meaning by the participants, and it was usually dependent on personal experience, role, and technical/non-technical competences. Generally understood as just or fair treatment for customers or students, it was also extended to comprehend fairness for the organisation or a more environmental reading which in some cases would have justified more intrusive forms of targeted marketing.

Similarly, accountability was understood differently according to the experience, competence, background, and positionality of the participants. For instance, accountability was linked to:

- Lack of transparency of algorithms by ML experts.
- DP breaches and staff accountability by the DPO.
- System stability and efficiency by those in technical roles.
- Training data by leaders.
- The relation and shared responsibilities between organisations and third parties.

All those interpretations of the concept are different aspects of it, and all of them are necessary to guarantee the general accountability of the organisation, as per GDPR.

Transparency was strongly linked to accountability. ML opacity was an issue, but low levels of transparency in decisions were viewed as an acceptable trade-off in contexts and situations where the possible consequences on Data Subjects were low. Such situations could be viewed as a practical example of the risk management approach offered by the GDPR. However, the criteria used to classify the risks for Data Subjects, or their eventual involvement, were unknown. Furthermore, the DPIA which permits the assessment of fairness in a new system did not appear to be used to satisfy that function, but was, rather, a bureaucratic necessity.

Another area was related to FAT principles and AI. The discussions around FAT and AI were usually focused on biases in data and algorithms. The debate around FAT principles and AI ethics in relation to the implementation and use of AI is still minimal. In general, some participants were able to appreciate the importance of discussing some aspects around FAT, but that was more around "traditional" technology and DP issues. For example, while different levels of staff accountability in managing personal data were discussed, personal responsibility in managing AI systems processing personal data was not. The management of AI was not generally treated or seen differently by most participants, nor was it seen as a potential source of unfairness.

Another area was related to the perception of AI and humans. It is quite interesting to see how the role of technology and the role of humans were perceived in relation to fairness and accountability. The technology is generally considered efficient and capable of delivering what is expected, without producing unfair consequences. The humans implementing and interacting with ML are generally considered able to correct potential malfunctioning, stopping the production of unfair consequences if they occur, a possibility rarely envisaged.

As seen in CS1, the original ML project was demoted as being too autonomous and risky for the accountability of the organisation. The human in the loop was considered safer, but no specific provisions were made for training and supporting the humans in their interactions with the machine. Similarly, the app in CS2 is expected to work without any issues, but when these arise, they are read as the result of a user's error (i.e., a potential candidate who misused the app) and not as a technical problem. In this case, too, the human (i.e., HR staff) is expected to intervene, correct manually the error, and proceed to validating the identity.

Therefore, AI is generally expected to work well, potential issues are often read as originating from users, and the human in the loop is generally assumed able to intervene and correct the problem.

Another area was related to changes in societal values. When societal values change, reflexive organisations adapt. Case Study 1 is emblematic. Leaders were aware of societal debates around gender and realised their self-identification criteria were inadequate to represent the complexity of students. The management adapted to that societal change, integrating it into the values of the organisation and its management practices. The multidisciplinary knowledge of senior management facilitated this process. The decision had an impact on the special categories of personal data collected from students, and it shows the capacity of the organisation to react to external changes, reflecting and acting to increase fairness and inclusion.

Therefore, the ways societal values were interpreted and included in organisational practices were multifaceted:

- AI was generally assumed to be faultless, and humans are supposed to be able to identify and correct system malfunctioning.
- Meanings and perceptions of principles can sometimes be taken for granted, impacting the execution of projects and their consequences for the organisations.
- Implementing AI projects that have ethical or responsible aims can provide too much reliance on general accountability and fairness.
- The management of AI does not always consider the peculiarities of AI.

- Changes in societal values can make the data for AI quickly outdated.

This demonstrates that strategic choices, different interpretations and meanings of the FAT principles, their specific relation to AI, the real capacities of AI and humans interacting with it, and the changes in societal values, all need to be accounted for.

d. Power and power dynamics

The power of organisations and subjects in this study can be described as multifaceted and performed in different ways. There was the power of internal subjects (associated with role and positionality), power created by new GDPR requirements and by organisational structures, discretional power of subjects, and power of the organisation in their relation to stakeholders. Analysing how power was played out is important for responsible management. For example, power and positionality. The dynamics of power in CS1 showed how the power of different subjects in organisations shapes the result of the technology. The role-reverse scenario presented to participants highlighted the power imbalance and the different degrees of protection of permanent staff, consultants, and students. The difference in power between staff and students is striking. The latter can be at the receiving end of the AI system, while the staff is predicted to resist a similar system measuring and predicting their performance, thanks to the opposition of trade unions. Considering the reactions of participants, the appearance of micropolitical forms of resistance would not be completely excluded in similar situations. Similarly, consultants (temporary technical staff) were also available to create similar systems for permanent staff but appeared to be less willing to be subjected to them. This also underlines the influence of external stakeholders in shaping responsible technologies, an aspect which deserves to be better explored. Similarly, a different perception of profiling and its necessity, or fairness of the practice, emerged quite clearly. Profiling was seen as a dangerous and unethical practice by a key participant (CS1), who also did not hold in high regard organisations making decisions on predicted behaviours. The same person, however, could not see any issue with the project, as students were going to be part of the institution only for a short time and they were not expected to be supported less in the case of a predicted low score. The fact the participant held a key DP role does not make this ambivalent view less problematic.

This case also shows the importance of involving Data Subjects' requests in responsible management. While the involvement of the users is advisable, the example above also demonstrates that, for responsible AI management, the presence and involvement of groups who can exercise collective bargaining power can also increase the protection of the rights of Data Subjects inside the organisations. This would create, internally, a similar provision made by Art. 80 of the GDPR that defines the rights of Data Subjects to give a mandate to representative bodies.

All these situations highlight how power was strictly linked to the positionality of the subjects, and vertical (staff-students) and horizontal (permanent-temporary staff) dynamics. Moreover, the perception of such dynamics was very much dependent on the context, influencing resistance or opposition to innovation practices.

Another example was the power created by GDPR and by the organisation. The GDPR created new obligations for organisations, for example, the new requirements around the DPO and the DPIAs, which need to be performed in order to achieve compliance. But such new obligations can also be read from another angle. For example, using the primary/secondary instrumentalization by Feenberg (in this case, applied to the process and not the technology), these new obligations that are "adapted" to a new context are also new capacities, and new forms of power, that organisations are given by the Regulation to perform new specific functions. How these new capacities are received and performed in those contexts, also taking into consideration sector and size, deserves to be explored with attention. The praxis of DPO, DPIA and effective power of managers identified in the findings are all interesting examples of the difference between the prescribed rules and their organisational practices.

For instance, the GDPR strengthened and extended the role of the DPO to the private sector. How this new obligation/power is performed varies across sectors, and it depends very much on where the DPO operates. Not always appreciated in the public sector due to its perceived power (as seen in 4.3.2), the DPO seems to be more appreciated in the private sector, but this often occurs within large organisations which have a low-risk appetite. Furthermore, it is not unusual for internal people to hold the role of DPO (e.g., CEO, IT Director, internal lawyers), a very problematic practice that cannot always guarantee the key independent function of the role. The GDPR also extended the provisions for the DPIAs. Their praxis showed how DPIAs were often considered a cost, rarely performed in the private sector and sometimes too much in the public, and not appreciated as useful management tools to evaluate fairness or the impact of innovation or used strategically as a preventive DP tool.

Another example of the difference that can occur between the power given to perform an action, and how that is effectively performed, was found in management practices. Due to the lack of specific knowledge on AI, some managers implementing AI projects were delegating the performance of some functions to others and signing off documents without completely understanding the implications. The effective power and responsibility of those managers are clearly open to question.

Another aspect was the power to diverge. For instance, other examples of power displayed within organisations were the power to create new AI categories and the power to disregard ML predictions. The capacity to reflect on external changes and the power to react to them, for example by identifying new categories for clustering data, has an impact on responsible management. Having the power to create or alter data categories, and include or exclude specific identities (CS1), can clearly affect the level of fairness in decisions made by the organisations.

Similarly, the power of humans to diverge from ML predictions was mentioned by several participants (CS1, CS2). Did those making decisions have the real power to diverge from existing categories and to create new ones (power prior prediction), or the power to diverge from the ML result (power post prediction)? How and the extent to which to diverge are also a matter of the real power the staff is provided with while performing those actions.

And yet, there is another level of analysis. Could staff also diverge from the "script" associated with their role and performativity? This aspect is more linked to the social pressure operating in that specific context, which would be potentially stronger on permanent staff. Such circumstances clearly have implications for responsible AI management. While many participants recognised the importance of discussing topics around AI, they also revealed how that was often avoided so as not to impact on innovation.

Another aspect was the power in dealing with stakeholders. The power displayed in dealing with internal and external stakeholders is another key power. As seen in many interviews and documents, external third parties in particular had a crucial role in creating and shaping the technology. The power in selecting, choosing, and negotiating a service is heavily impacted by the amount of competition. If an AI service is offered by a few third parties displaying rather questionable ethical practices, the effective power and agency of the organisation in creating responsible innovation would likely be impacted.

This demonstrates that differential expression of power, such as power and positionality, power created by GDPR and by the organisation, power to diverge and power in dealing with stakeholders, all need to be considered.

e. Risks

Another critical aspect is related to the identification, perception, and assessment of risks. For example, specific risks linked to AI/ML. The interviews with the ML experts provided information around some of the specific risks of ML. The necessity to find a balance between past and new data, the need for retraining on real data to maintain performance and overtime efficiency, and the risk of a drop in efficiency when algorithms are used in the real world, were some of the risks mentioned. They can impact on the processing of personal data. The identification and assessment of such risks require specific knowledge, awareness, and frequent checks in order to understand when the algorithm is not working as expected, to raise concerns, and make/ask for corrections.

Similar provisions are also important in Augmented AI systems. Both organisations were planning to use such a model, as the chosen model by CS1, and as a fallback option by CS2 if the automated identity validation failed. Furthermore, in CS2 the use of ML is performed by an outside partner. The issue regarding the stability of ML is completely external to the area of action of the start-up, and when functionality issues occur with face recognition, they are read as resulting from a user's mistake. Therefore, the responsibility for similar mistakes gets lost in the processing performed by third parties.

Noticeable was the fact that neither organisation was considering or questioning the specific ML risks around algorithmic performance.

Another aspect was related to the link between risks, responsibility, and accountability. The lack of clarity around data, processes and relationships with stakeholders was frequently resulting from unclear responsibilities between various subjects/teams/entities (the problem of many hands). In similar situations, the responsibility of processing personal data was often perceived as lying with someone else (e.g., vendors, cloud providers, other departments and roles). In such cases, perceptions or assumptions can be more important than real meanings. For example, if temporary staff/consultants expect other roles to deal with fairness (CS1), or if the managers assume the algorithms created by others to be inherently fair (CS2), what is fair or not in the system being implemented or in its results can become secondary. If the problem is perceived somewhere else, it will be more difficult in terms of applying critical thinking or looking for second opinions or for assessing negative outcomes.

The GDPR tries to clarify the external relationships between different entities processing the data (e.g., controllers/processor/joint controller). However, the data exchanges that happen internally, between different teams or areas, are not subjected to any detailed requirement. Obligations or suggestions on how to regulate accountability in internal exchanges of personal data are missing in the Regulation. The use of AI adds further complexity. For example, the DPO expressed high confidence in the internal mechanisms assumed to guarantee the accountability of various internal subjects who process personal data. But the knowledge the same participant had around AI or around the project being implemented was rather low. Furthermore, their intervention was marginal in the DPIA.

Another aspect was related to risks and security. Most of the participants without ML backgrounds considered the ML risks a matter of cybersecurity, which could be strengthened by increasing the protection from external hacks. The security risks were seen as external risks, coming from the outside. ML was in general not seen as necessitating more attention than other technologies or bearing more risk for the protection of the personal data, which also transpired in the lack of specific staff training (CS1).

Therefore, the specific risks linked to ML algorithms, those resulting from unclear responsibilities, internal processes and security, and those risks resulting from the interaction of DP and AI, were mostly not considered. This demonstrates that different kinds of risks, such as specific risks related to AI, risks linked to roles, responsibility and accountability, and specific risks linked to security all need to be considered.

f. Processes, stakeholders, and decision-making

The main focus of the AI innovation cycle was project delivery. The strategic phase prior to the project and the post-project phases (BAU and long-term sustainability) appeared to be less present in the general discourse. Similarly, the discussion around DP was very general, related to the project, and not to strategy, use, or long-term sustainability of the system.

Assessments were seen more as obligations and necessary steps than important tools for innovation. Algorithm assessments were not performed or planned (experts, CS1), or not mentioned while discussing the processing performed by partners (CS1, CS2). Even though equality assessments were performed in CS1, they were not extended to the AI systems.

DPIAs were performed by one/two roles, right before the end of the implementation and when strictly necessary. No update of existing processes and documents was planned, nor were audits or long-term sustainability of the technical systems. Noticeable was the level of reflexivity within CS1, and how they decided to modify the model and the processes in response to a perceived high-risk of automated AI.

The relationship with internal and external stakeholders varied considerably. CS1 appears more cautious after the experience with the first vendor, leading it to choose an in-house system created with the support of technical consultants. It was unclear if and how the experience with the original vendor had any impact on selecting and dealing with the consultants. The level of scrutiny of competences of the consultants was unknown. CS2 displayed an interesting mixed approach to stakeholders, with a different level of risk attached. It was very cautious about the partners using their services via the data trust but completely comfortable with practices, reputation and processing performed by the data traders. The evaluations and checks adopted by CS2 in selecting the traders were unknown, and previous collaborations might have informed the decision. However, such previous services appeared to be more limited and used in different contexts where high levels of scrutiny were in place.

Contract details around controllers and processors and responsibility around decisions made on top of AI predictions were unknown.

Decision-making processes were also diverse. Many specific details were not shared with the researcher, and how decisions were made, or conflict managed within the Board or other roles, is not known. And yet, some information emerged. Both organisations had been working on their projects for a long time. They are different sizes and have different business models and AI could impact differently on their activities. CS1 had a more structured and regulated decision-making process, with more roles involved. However, some decisions appeared to have been made without the involvement of key internal stakeholders, e.g., during the selection and purchase of the first AI product. Similarly, decisions around DP appeared at times based more on an assumed level of risk regarding data and technology than an assessed one. The decisions made on top of predictions appeared to be completely dependent on the evaluations made by the user of ML. The decision around the potential communication of the prediction to the Data Subject/student was still being debated. CS2 had a more flexible structure and decision-making process. it was investing time in becoming better informed around AI, its risks, their level of accountability according to the GDPR, and in improving their DPIAs. And yet, it was noticeable how their concerns and decisions appeared to be more focused on specific aspects, e.g., some partners or some elements of the GDPR (e.g., consent), and not on others. While some concerns emerged around the ethics of the practices of some partners (such as the possible use of personal data linked to an adult website), those did not lead to any specific decision and action, such as assessing the quality of data, informing Data Subjects, or excluding similar special categories of data from their validations.

Therefore, the AI innovation praxis was mainly limited to the project phase. The assessments were generally experienced as an obligation and not an opportunity. The silo mentality impacted on the exchange of information, and on the adaptation of existing processes and resources to the new AI systems. Relations with stakeholders were shaped by the specific knowledge of people, their contexts, and past and current experience. Key stakeholders were often not included in the praxis. Decision-making reflected similar patterns and appeared to be dependent on self-reflexivity processes. Some participants displayed a certain level of awareness around the impact that their decisions and those of their partners could have on responsible practices and the rights and freedom of Data Subjects. And yet, some key stakeholders who could have supported the decision-making activities were not included in the process.

Thus, the above demonstrates that different elements, such as processes, how assessments are performed, how stakeholders are involved and how dilemmas are solved, all need to be considered.

g. Summary

The elements identified as important using the theoretical lens were various, for instance: strategic choices, diversity of meanings and understandings of FAT, specificity of the FAT principles in relation to AI, different understanding and assumptions around AI and humans interacting with the machine, changes in societal values, power and positionality, power created by GDPR and by the organisation, power to diverge, power in dealing with stakeholders, specific risks related to AI, specific risks linked to security, risks linked to roles, responsibility, and accountability, processes and assessments, stakeholder management.

Table 7-2 Elements identified using the CRAIDA framework

• Strategic choices and decisions	• Diversity of meanings and understandings of FAT	• Specificity of the FAT principles in relation to AI	• Different understanding and assumptions around AI and humans interacting with the machine
• Changes in societal values	• Power created by GDPR and by the organisation	• Power to diverge	• Power in dealing with stakeholders
• Power and positionality	• Specific risks related to AI	• Risks linked to roles, responsibility, and accountability	• Processes and assessments
• Stakeholder management			

In the next chapter, the model will be presented, illustrating how these elements have informed the model for responsible AI and DP management..

7.7 Developing *CRAIDA*: theorisation, research process and sequential analysis

Developing a suitable theoretical framework for the research was a long journey. The framework presented in this chapter was the result of a long process drawing upon the reflection on empirical and theoretical research, fieldwork, analysis, and an engagement with philosophical questions. This was a continuous process during the course of the whole research.

The literature review provided the researcher with a sound understanding of the significant research areas. At the same time, the reading done around different theoretical positions also created a sound basis regarding the understanding of the critical tradition in theorising technology. However, no ready-made existing positions appeared to be suitable for theorising the complexity of the interplay between AI, DP, FAT, and the agents within certain organisational contexts.

This framework – as it has been described in this chapter – was also significantly shaped by an initial analysis of the research data, plus an ongoing quest for nuanced theorisation. The analysis of the data from the interviews and the case studies was carried out considering the key areas (AI, GDPR, FAT), and some key factors identified in the literature, such as use of the technology, understanding of GDPR, specificities of AI and DP compliance, and different perception of risks. The elements identified within this analysis were expected to be important for answering the research question. A continuous rethinking of key concepts from within RRI first and CTT – in confrontation with the empirical insights gained through analysis – slowly provided new theoretical tools for sharpening the focus. New questions emerged around the role played by the people in those contexts, their active role in shaping processes, and the outcomes resulting from a different praxis.

This allowed for a deeper level of analysis, in which both traditions are facilitating the understanding of complex of external and internal factors. While the 'first' data analysis helped explore how such organisational realities were constructed, the 'later' analysis carried out using the *CRAIDA* framework permitted the researcher to see how the dynamics of power operating in such contexts shaped those praxes. By critically (re)reading the results of the empirical data, a deeper level of analysis helped explain the conditions under which decisions were made. The refined theoretical lens provided by *CRAIDA* permitted, for example, the researcher to identify the multidimensional and complex nature of power operating in those contexts, and the composite role and power of different entities operating in multiple ecosystems. Such a deeper reading would not have been possible without going through a long reflexive process which developed with the research. It made sense to present, capture, and fully validate this level of analysis in a separate chapter. This also allowed the researcher to develop a more widely applicable theoretical tool (*CRAIDA*), which in turn facilitated the creation of two models on responsible management described in Chapter 8 of this thesis.

7.8 Conclusion

This chapter presented *CRAIDA*, the new theoretical framework developed to critically read organisational contexts. The chapter firstly introduced RRI, with its main characteristics, and its relation to privacy, DP, the industry, and its most recent developments. It then presented some critical approaches within critical theory, CTT and Feenberg, and his work on the role of the subject in relation to technology, power, and values. The chapter then focused on *CRAIDA*, the critical AI&DP Management framework inspired by RRI and CTT. After presenting the main elements of the framework, the chapter gave a detailed analysis of the aspects identified through the theoretical lens in the research data, such as issues around decision-making, power, and risks. Such factors impacted on and shaped the implementation of AI in those contexts. The chapter proceeds with some further reflections on the theoretical foundation and the development process of *CRAIDA* as a theoretical model.

The insights from the analysis proved essential for the creation of models aimed at supporting management in implementing responsible AI and DP innovation (Chapter 8).

CHAPTER 8: A PROPOSAL FOR RESPONSIBLE AI AND DP MANAGEMENT – THE *RAIDIS* MODELS

8.1 Introduction

This chapter presents a proposal for a responsible AI and DP management model informed by the findings of the research. Firstly, the chapter introduces the Information System (IS) framework as an analytical tool for a responsible AI&DP IS/*RAIDIS* management model and explains its five key factors. Next to the *RAIDIS* management model, the chapter further presents the maturity model, *RAIDIS MM*, that shows the various stages towards a responsible organisational strategy. The chapter then ends with a synthesis of the *RAIDIS* models.

8.2 Using the IS framework as an analytical tool for developing a model for responsible AI and DP management

There is no scarcity of guidelines published by institutions and organisations which detail what organisations are expected to do in order to work with AI ethically (as seen in 2.4.1). While being well intended, these guidelines have often focused on data and do not consider the challenges organisations face and their specific contexts. While identifying and correcting biases embedded in data is crucial for increasing responsible innovation, this is not sufficient. Data is only one element of a complex environment where people make continuous choices and decisions which occur in different moments within the innovation process. Additionally, as seen in Chapter 7, while existing guidance encompasses suggestions on what to do, it often lacks specific suggestions on how to do it (Ryan & Stahl, 2020). Therefore, responsible AI and DP management necessitates a different approach. How Stahl sees privacy as related to other aspects of RRI (e.g., activities, actors) (Stahl, 2012) and his view of RRI as a meta-responsibility related to existing responsibilities (and not as a new type of responsibility) or existing processes, was inspiring for the creation of the model presented in this chapter.

Thus, this research proposes a new approach that encompasses different elements present in the environments and promotes responsible practices starting from what is already done in such contexts, for example, processes performed to comply with GDPR requirements. Such an approach is not merely focused on a technical point of view. Technology and data are part of socio-technical systems which bring to the fore the role of human actors in those environments. Consequently, this research considers human decisions at the core of responsible innovation. Leaders and managers have to choose between different options, trade-offs, or dilemmas (e.g., more autonomy vs less control, more data vs privacy). Human decisions do not happen in a vacuum of values. They are often taken rapidly and made in environments where different concomitant factors all play a part. Roles, competences, power, experiences, backgrounds, personal ethics, personal and organisational self-reflexivity, internal regulation and culture, and technical capabilities all contribute to decisions made and innovation created in those contexts. Thus, the path towards responsible AI and DP management is more complex than the one suggested by current AI ethics studies, which are mainly focused on unbiasing data. Fairness, accountability, and transparency are experienced by different subjects throughout the innovation process. Additionally, those making decisions have also to explain to various subjects (regulators, stakeholders, customers) why and how decisions were made.

The researcher considers a framework based on Information Systems (IS) management to be extremely useful in exploring such complexity and unveiling how technology, people, and processes are interconnected. As IS can provide a useful framework in the implementation and compliance of DP/GDPR projects, a similar framework can prove extremely valuable in the implementation of AI. Yet, the organisations interested in implementing responsible AI projects need a more refined IS model, inclusive of different elements, inside and outside the organisations, having the capacity to impact on the innovation milieu.

The following section presents an enhanced IS model for responsible AI and DP management that provides organisations with specific tools for managing the implementation of AI systems. By identifying the exact risks and potential within specific areas, it creates the foundation for responsible AI management that strengthens innovation and DP compliance. The model complements existing general guidance by providing specific suggestions of what needs to be put in place and how to enable responsible management.

8.3 How the research informed the model

The model was created considering the key points identified via the data analysis (Experts, CS1 and CS2) and the new theoretical framework (Chapter 7). In the following, the points are introduced considering the two different origins.

Table 8-1 Main points identified via the data analysis

Synthesis main points (SURVEY+CS1+CS2)	
Technology • Fast innovation • External pressure (market) • AI powerful technology (but not perfect) • Knowledge and perceptions of resources (AI, data, people) • Resources (inside outside) • Technological self-reflexivity (capacity of the organization to understand the impact of its technology)	**People** • Information/knowledge in different people • Information/knowledge exchange • Org. structure, roles, responsibility, and link to accountability and assumptions • Relationship of AI and people. Human supervision of AI • Empowerment of staff and end-users • Different understanding of AI, GDPR and FAT • Diversity in experience and backgrounds
Processes • Importance of the full innovation process • Internal processes and external processes (the first reducing, the latter increasing) • Holistic approach to risks • Various assessments and innovation tools • Augmented AI processes • Importance of clear rules	**Stakeholders** • Stakeholder identification and management • Source of expertise and/or risks • Extremely fast innovation created by start-ups • Potential loss of control and hacking risks • Complex external ecosystem of third parties providing specific services (e.g., data brokers)

Decision-making • Decision-making around strategy, project, and models (automated and augmented AI) • Business model and decision-making • AI and business model (technology often created before the business case)	**Innovation** • Fast innovation and reduced time for assessing it • GDPR good/bad for innovation according to context and awareness
Power • Linked to positionality of stakeholders • Impact of power dynamics on implementation, use and end-users	

Table 8-2 Main points identified via the theoretical approach

Synthesis main points (*CRAIDA*)
Strategic choices, decisions, and dilemmas
Diversity of meanings and understandings of FAT
Different understanding and assumptions around AI and humans interacting with AI
Specificity of the FAT principles in relation to AI
Impact of changes in societal values

Power: power and positionality, power created by GDPR and by the organisation, power to diverge, and power in dealing with stakeholders
Specific risks related to AI, specific risks linked to security, risks linked to roles, responsibility, and accountability
Processes and assessments
Stakeholder management

All the points above, with the knowledge gained via the literature review, have informed the thinking of the researcher. The model is the result of such processes, and it is presented in the following section.

A cautionary remark is needed in order to reduce some expectations. The elements of the model are usually the results of a combination of multiple factors identified by the researcher in the data. For instance, the Augmented AI process presented later in the chapter is resulting from the reflections made in different moments. For example:

- While reading on ML predictions, specifically the work by Agrawal et al (2018) on prediction machines.
- While thinking about the factors identified in the data. E.g., how organisations were choosing and using AI to augment decisions (experts, CS1, CS2) and imagining the use of the technology and personal data in the future.
- While reflecting on some *CRAIDA* points (e.g., decision-making and power).

Such complexity is typical of almost every element of the model. Thus, trying to identify only one source for such elements would be inaccurate, and would undermine the complexity of the identified phenomena.

8.4 Key elements of the *RAIDIS* management model

The most significant areas to consider for a model providing guidance on responsible AI and DP management include the three key IS elements: technology, people, processes.

Thus, the complexity of organisational practices identified via the analysis of the data highlighted the critical relevance of another two elements: stakeholder management and decision-making processes. The relation with external and internal stakeholders proved to be a key factor in supporting organisations or exposing them. Similarly, decision-making processes were revealed to be crucial for responsible practices. Thus, in order to recommend a model for responsible management that addresses the complexity of organisational practices, five elements should be considered: technology, people, processes, stakeholders, and decision-making. The following discussion aims at illustrating how these five key elements and their interrelations can impact compliance and responsible innovation.

The following model mentions DP when referring to DP in general (not a specific DP regime), and the GDPR when precise requirements of the Regulation are discussed. Similarly, the text mentions AI when referring to the technology in general, and ML when it refers to ML's precise characteristics (e.g., prediction).

8.4.1 Technology

In the following, the important elements around technology and data are presented.

Data and biases. The conversations with the participants and the general debate around AI ethics tend to be mainly focused on data and biases, mostly input or training data. However, all types of personal data processed by organisations should be carefully considered and assessed for potential issues impacting FAT and compliance, not only input data and training data (used to train the algorithms), but also output data (produced by the application) and feedback data (resulting from the environment where ML is deployed).

Particular attention should be given to special categories of data, such as biometrics (e.g., bodies, faces, voices, health data) used to classify people, verify identities, and provide services. Issues can originate from the use of more training data from a specific category (e.g., men), body features can be misinterpreted (e.g., misgendering trans people), and people can be denied services or suffer serious consequences (e.g., black skin not being recognised). The lawful bases and the scientific validity of AI systems that process biometrics should be assessed considering the potential for higher impact on the rights and freedom of Data Subjects and the increased risk for non-compliance and reputational damage. Thus, organisations should not only make sure their systems are not the result of, are not using, or are not producing biased data, but also that such applications are not based on questionable scientific evidence (e.g., Emotion AI).

Furthermore, the findings of this research reveal that data is not the only element capable of causing some ethical issues. Organisational practices can also bear some risks (see other factors below), and these also need to be identified, addressed, and corrected.

Therefore, data, systems and practices can all impact on responsible management, and their impact on different categories of users should be considered and assessed.

Quantity, quality and variety. Particular attention should be given to the quantity and quality of data. Increasing the quantity of personal data does not necessarily increase the quality of the information on Data Subjects. It does, however, increase the risks for DP and Fairness. Furthermore, whereas data is homogenous, increasing its quantity does not make up for its lack of variety. As group thinking is to be avoided within organisations, this should also be done with data. Diversity and variety of datasets bring different perspectives and limit the stereotyping done with ML.

What ML does better. The organisations choosing ML should understand what algorithms can do well. Algorithms lack flexibility and are unsuitable for soft targets. While they are good at identifying patterns, the meaning of such patterns must be carefully interpreted (e.g., correlations vs causations). While algorithms are good at making predictions based on past occurrences, they can also reproduce past unfair practices. For example, the prediction of criminal recidivism could also predict the re-arrest resulting from biased police practices. While algorithms can make more accurate predictions, they cannot predict the future. While they are good at making accurate predictions with clear past data, the same does not happen when past conditions have changed, or past data is not available. In similar situations, humans can make better predictions.

Behavioural data. Human behaviour is the result of many factors, and not all factors which impact the human experience can be measured. This impossibility makes the predictions of human behaviour extremely difficult to achieve (if not impossible). Organisations should be aware of this. If an organisation decides to implement systems aiming at predicting possible future behaviours, it would be advisable to consider not only past behaviour, but also new factors that in that context might be impacting negatively on the behaviour. For example, if some Data Subjects are not doing something done in the past, due to a new factor, or if they have never performed a specific action due to the peculiarity of that context (e.g., institutional racism), such factors should be considered.

Thus, organisations interested in predicting people's behaviours should be aware of the limitations of such practice, and not only look at what was done and how, but also at what caused that behaviour in the past (why).

Opaque algorithms, complexity, and necessary checks. Organisations using "opaque"/black-box algorithms, such as deep learning, should pay particular attention to their development, outputs, and context. As per GDPR, organisations are responsible and remain accountable for their processing, and must be able to explain and back up their decisions. As already happening in many organisations, the complexity of ML systems often becomes one of the reasons why management avoid asking stakeholders (e.g., developers/vendors) some specific information about the system.

Conversely, the potential complexity of ML cannot be an excuse to justify a lack of accountability. Managers need to know how the system was created, and which checks and security measures were considered. They should know, for example:

- the types of data used for training the system and how they are used in that context.
- if there is any loss of accuracy when ML is used in the real world.
- how often the algorithms should be retrained and the accuracy of their predictions.
- the defence measures embedded in the system against external attacks.
- whether the algorithm is transparent and auditable.

When the system is acquired, organisations should request and assess specific documentation. For example, information on AI systems previously created by external consultants, and DPIA, technology and equality assessments (if performed) should be requested from vendors/third parties selling the technology. When the system is implemented, organisations should plan how the algorithms will be continuously tested, updated, and audited, how often, and by using what kinds of data (ongoing performance in BAU).

Knowing the context. The context is an important factor for FAT. Knowing the context in which the system will be used increases the accuracy of the selection and the quantity of data to be used.

Specific attention is also needed for data collected from highly changeable contexts. Rapid socio-economic and political changes can impact greatly on the environment and individuals. The volatility of the data collected from such contexts should always be considered when that data is used to train ML, as this can avoid using data that is already unsuitable.

Many systems are created by engineers who do not know where they will be deployed and are implemented by managers who do not fully understand them. Both situations can create unfair outcomes and accountability issues.

Security. A wide understanding of security issues is advisable, inclusive of physical security, technical security (e.g., internal access) and cybersecurity. Security should aim at protecting the data against external access with the intent to steal the data, alter the integrity of the information, and alter the availability of data.

Transparency around the use of AI. Organisations should be transparent about their use of AI. Data Subjects should know when AI is used for processing, how it is being used, which data is processed, its granularity and how the decisions resulting from automated or augmented processes can be challenged and redressed. Thus, transparency should not be limited to input and training data but also extended to the system, output, and feedback data.

Organisations should be able to request and provide similar information to current and potential stakeholders.

8.4.2 People

In the following, the important elements around organisational structure and key competences are presented.

Responsible AI Officer (RAO). The roles and competencies of IT directors, project owners, IT project managers, Information Governance managers, privacy directors and other potential roles (such as ethics managers/officers) within organisations do not always include the wide and holistic understanding that is needed to manage the full innovation process of AI in organisations. For this reason, organisations should appoint an RAO in charge of the full AI innovation cycle. This new role would require knowledge and expertise around:

- AI: the most recent research and applications of AI technologies, with a sound understanding of the internal and external impact of innovation.

- DP: the most relevant legislation, both nationally and internationally, and an understanding of the politics of DP, inclusive of the geopolitical aspects.
- Ethics and responsible governance: not limited to the ethics of data but inclusive of a deep understanding of current debates around innovation, socio-economic and political issues, ethics, power, and intersectionality in specific contexts.

The RAO should have effective authority and power, be part of the senior management and not be an independent officer as is the DPO.

Responsible environment and effective participation. Responsible management is an organisation-wide effort. Every permanent and temporary role should be part of (and feel part of) a responsible working environment. The GDPR aims to create an environment where DP is everybody's concern and responsibility. Similarly, AI affects various roles, not only those in charge of IT. Likewise, FAT and responsible innovation should not only be a concern for specific roles and should not only be enabled via a top-down approach, which could be resisted by staff. This participation should be part of the innovative creative milieu from the beginning. Participation in knowledge exchange can reduce micropolitical forms of resistance. When staff are part of the process, are informed, understand the system, trust it, and acquire new skills, they then can use the technology, participate in shaping innovation, be less concerned around job losses due to automation, and become an asset for the organisation.

Training. Training should be specifically tailored according to industry and context, and consider characteristics, vision, aims and challenges of the organisation.

Specific training on AI and DP would be advisable for leaders and managers. When ML is employed, training should include information around the specific characteristics of ML, capabilities, and risks linked to the processing of personal data. Training is also necessary for staff. Training on both AI and DP should not be conducted solely online; managers need to exchange knowledge with staff and ensure they acquire an accurate understanding of the topics. Such learning activities could be included in a more general strategy aiming at creating spaces where individuals can acquire and exchange new information and have a more active part in innovation. In such spaces, ideas, doubts, and perceptions could be challenged and shared.

Organisations implementing Augmented AI should create specific AI-human interaction training aimed at identifying and tackling technical and human potential issues. For example, errors and malfunctioning in algorithms and over-reliance on predictions.

Diversity and innovation. Organisations using AI should be particularly careful about the risks posed by group thinking. Diversity is necessary to reduce bias in data, models and processes, and diversity of opinions, backgrounds and expertise should be encouraged and actively sustained by management.

Roles and responsibilities. Roles and responsibilities of staff should be clear. Several factors could alter the necessary exchange of information and expertise, creating issues for responsible innovation. These issues can occur in three different areas.

1. Responsibilities of stakeholders. The requirements, limitations, and boundaries of the assignments of external stakeholders should be clearly defined.

   The expertise and experience of temporary staff (e.g., developers creating the system) are precious resources for organisations. This is particularly important for organisations implementing emerging technologies projects. Managements should make sure that the knowledge of consultants is shared, and that no misinterpretation of the GDPR limits or blocks such exchange. The potential risk of high fines in the case of wrong advice could be seen by some consultants as hardly justifying the desire to provide information around areas outside their specific assessments. Such withholding of information could ultimately impact on their innovation.

   The creation of some free spaces (real/virtual), where different opinions and information could be shared and where different staff members can safely interact, could foster a diverse and multidisciplinary environment in organisations interested in responsible innovation.

2. The balance between technology and other disciplines. Different roles within project teams can have a different perception of risks. Technical roles can be more focused on the technical efficiency of the system, while other roles could be more attentive to factors impacting societal values. Thus, while creating environments where different disciplines have their voice, specific attention should be given to the risk of creating contexts where prominent tech bubbles or other areas become too influential, impacting upon a different point of view and ultimately on innovation.

3. Expectations around specific roles and expertise. The specific contribution that particular roles can bring to the project was another aspect that emerged from the findings. The SMEs in the case studies were involved in specific moments and expected to contribute around their areas of competence. While this is a common practice in project management and it is perfectly understandable, the same experts could also contribute to debates happening in other moments, for example, while defining risks or discussing issues around FAT. Such internal barriers due to fixed processes and expectations linked to roles and competences can impact innovation. If data scientists are usually not expected to discuss issues around ethics and fairness, and their involvement in projects typically occurs when such discussions have already taken place, this can be counterproductive for responsible innovation.

   While the general debate on AI has focused in recent years on the need to train data scientists around ethics, the modalities of their involvement in organisational contexts are mainly under-explored. Altering the moment and the content of their expected interventions can contribute to shifting assumptions and expectations. This could lead to widening and diversifying dialogues around responsible innovation which go beyond the expectations linked to the performativity of the roles.

A system-wide approach to AI and DP, inclusive, diverse, multidisciplinary, and based on exchange and dialogue can foster more responsible environments. As noted by Feenberg (2010), how subjects understand the technology is crucial. Via a secondary instrumentalization, the object is adapted to a specific context thanks to active participation of the subjects. Also important was his focus on the democratic transformation from below (Feenberg, 2005) which can contribute to an "enhanced", more innovative, socially acceptable, and desirable AI.

However, the definitions and understanding of values should not be taken for granted but clarified using concrete examples. For instance, people can have different understandings and perceptions of fairness, or what responsible innovation means. This can also happen within the same organisation or team. Fairness can be linked to justice, ethics, or to the expectation that data is going to be processed in a certain way. Some leaders might also understand the concept as more linked to the sustainability of the organisation, while others might link it to societal impacts.

Thus, the same understanding of fairness should not be taken for granted but be made clear in the whole innovation process while defining the strategy, delivering the project, using the technology, and evaluating the long-term sustainable impact of projects.

8.4.3 Processes

Organisations that want to establish fair, accountable and transparent AI practices need to carefully consider and assess internal and external processes. In the following, the important elements around processes are presented.

Which processes? The processes performed at every stage of the full innovation cycle should be assessed. For example, the processes performed while:

- Defining the strategy.
- Ideating the project.
- Acquiring/developing AI.
- Completing the project.
- Using the system in BAU.
- Evaluating the impact of AI both within the organisation and externally.

Assessing risks. The risks linked to the use of AI should be assessed with the help of internal stakeholders and experts consulted before acquiring AI systems. Buying or building AI systems involves different risks to be evaluated via mandatory, and in some cases advisable, assessments.

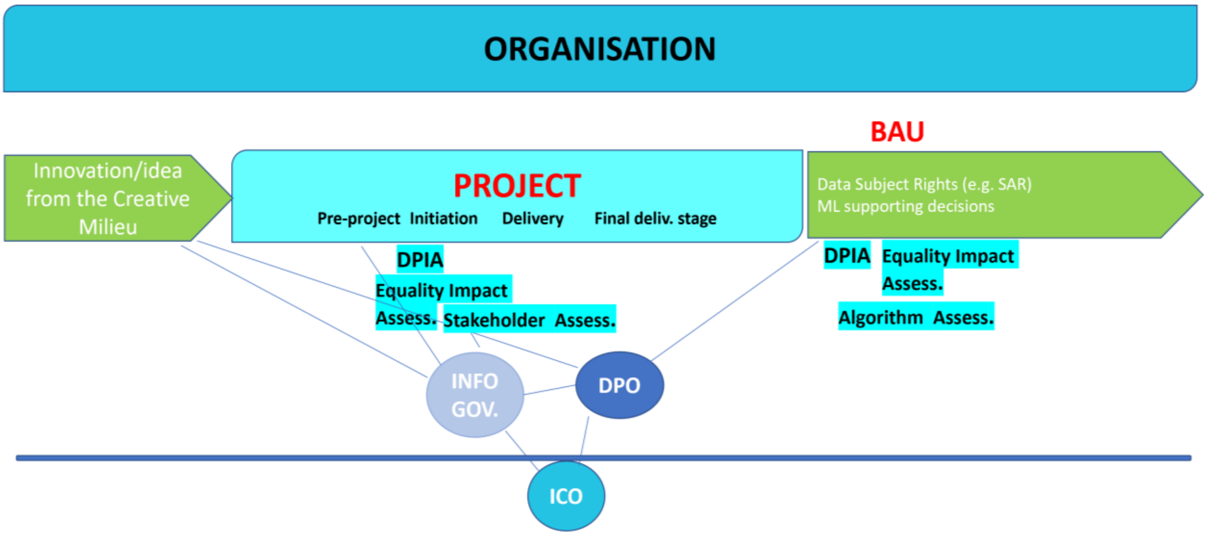
The DPIA is the most important assessment. It should be seen as a precious tool and not a box-ticking exercise. It makes it possible for different competences, expertise, and teams to come together and to evaluate the risk associated with that use of AI. When the skills needed for performing DPIA are not present within the organisation, these should be acquired externally. Data Subjects should be heard as part of the assessment.

DPIA should be performed at the early stages of the project development (not at the end of the initiation stage), after the creation of the business case (Figure 8-1). The organisation should perform DPIAs before investing resources on AI projects, as these could be vetoed or blocked by the roles in charge of DP approving DPIAs. The DPO should be involved during the assessment. Similarly, the involvement of the ICO is required as soon as the awareness of the residual risk (unmanageable by the organisation) emerges, and not just before starting the processing.

Other assessments. Algorithm Assessments and Equality Assessments are two specific assessments to be performed with the support of precise expertise and completed before the processing. Performing such assessments in BAU would also be advisable.

While performing DPIAs is generally required for AI systems, performing privacy impact assessments is not required. However, organisations should also assess the risks for privacy, considering the power of AI to merge information from different sources, and the increased risk of the identification of Data Subjects following the aggregation of non-personal data. While such data are outside the GDPR scope, performing such assessment would increase risk awareness and accountability, and would reduce the risk of unexpected GDPR fines.

Figure 8-1 The AI and GDPR innovation cycle



(Design: Chiara Addis)

Assessing stakeholder risks. When vendors or consultants are involved in selling or building the system, the organisation should request the necessary information to assess the risks associated with those partnerships. Similarly, the risks posed by the processing performed by internal stakeholders should not be underestimated but assessed, as careless internal practices can also impact on the final innovation outcome.

Assessing the understanding of fairness. As different people can have a different understanding of the concept, their understanding and perception of fairness should be made clear.

BAU. The potential impact of AI systems on the rights of Data Subjects should not stop with the implementation of the project but be continuously assessed in BAU. The assessments should not be considered as one-off documents, but be periodically performed and reviewed, and their completion made mandatory. This would be especially important in the case of ML algorithms (which learn continuously), new data feeds added at a later stage, and when an already existing system is used for different purposes.

Creating/reviewing documentation. New documents and privacy policies should be created or reviewed in order to take AI into account. For example:

- Different DPIA templates should be created considering AI/ML and the level of technical knowledge within different areas of the organisations.
- Clear information around the level of granularity of data should be included in the privacy policy.
- Clear information on how decisions are formulated in an Augmented system should be communicated to users (e.g., prediction + human decision).

The review of documents should also be intended in restrictive terms if, for example, the assessments of existing clauses would reveal the risk of misuses with new AI systems. If a clause can be interpreted broadly, justifying processing that can create unfair consequences, such a clause should be restricted (preventive DP).

Sustainability. Future envisaged processes regarding the long-term sustainability of systems should encompass all resources (human, technological and organisational).

8.4.4 Stakeholders

Organisations that want to establish responsible practices need to carefully manage internal and external stakeholders. In the following, the important elements around stakeholders are presented.

a. External stakeholders

The GDPR regulates the relationship between controllers, joint controllers, and processors. Specific attention should be given to vendors selling AI and to third parties processing personal data on behalf of controllers and processors.

Those organisations investing in responsible innovation and AI ethics should pay particular attention to the management of external stakeholders.

A growing number of companies are using AI to provide services to organisations. The ethics of some of such practices are often open to question (as seen in 1.2.1). Thus, high due diligence is needed around the reliability of stakeholders.

Stakeholders offering such services should provide some information around the technology used and their FAT practices, for example:

- Security and organisational measures.
- Sources of the data used to train their algorithms (e.g., data from the Internet or other countries).
- Data quality control: DPIA, algorithm and equality assessments.
- Specific information on algorithms (e.g., stability, accuracy, and retraining).
- Specific information around past projects.

Such information can help the organisation to understand the reliability, competence and knowledge of the party providing a service for whom the organisation is accountable.

The general readiness and openness of partners around their practices is another important factor. The existence of trade secrets and confidential information used to deny information should be carefully evaluated. Future rules that can help to clarify what can be safely disclosed would be important considering how often such rules are used to conceal questionable practices.

Contracts with third parties should be clear around data processing, competences, and responsibility. Data should be deleted after use in order to protect Data Subjects from the risk of future re-identification. Such deletion would also prevent potential undetected biases impacting on other systems when used for training.

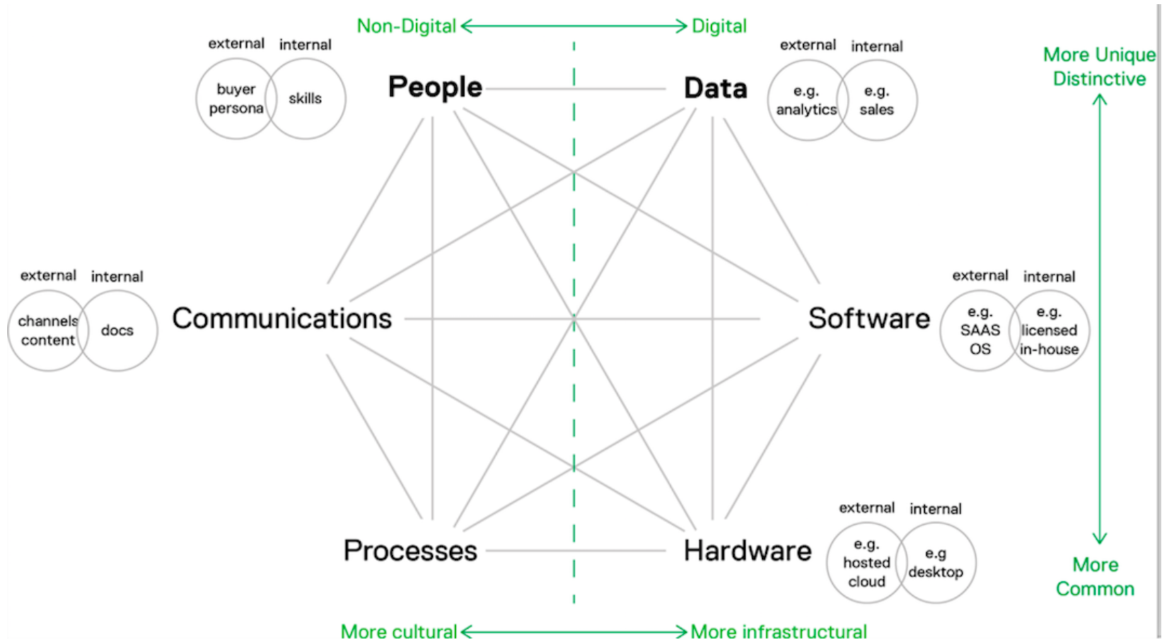
Boundaries and responsibilities should be well-defined to prevent issues similar to the internal *problem of many hands* that could occur with external stakeholders. This is particularly relevant for accountability and multiparty liability in multiparty ecosystems.

Specific clauses should also be precise regarding the success criteria used to evaluate the system: for example, efficiency and/or potential impact on societal values (e.g., discriminatory effects). Attention should be given to the language used with stakeholders, considering some vendors can be more used to sales and technical terminology than ethics.

The experience and competence of vendors and consultants are precious resources for organisations. However, stakeholders can also be a liability. Thus, the use of such external entities processing personal data provided by the organisation needs to be carefully managed. Furthermore, the relationships between external third parties often lack clarity around accountability, data, and the purpose of processing. If a third party is using other partners to create or process that personal data, the organisation needs to be aware of who is processing what, why and possibly how.

Therefore, while digital organisations can appear to have simpler organisational structures and internal processes (Figure 8-2), such considerations should always be carefully assessed considering the growing market of entities providing specific services.

Figure 8-2 The Digital Business as an Information System Model

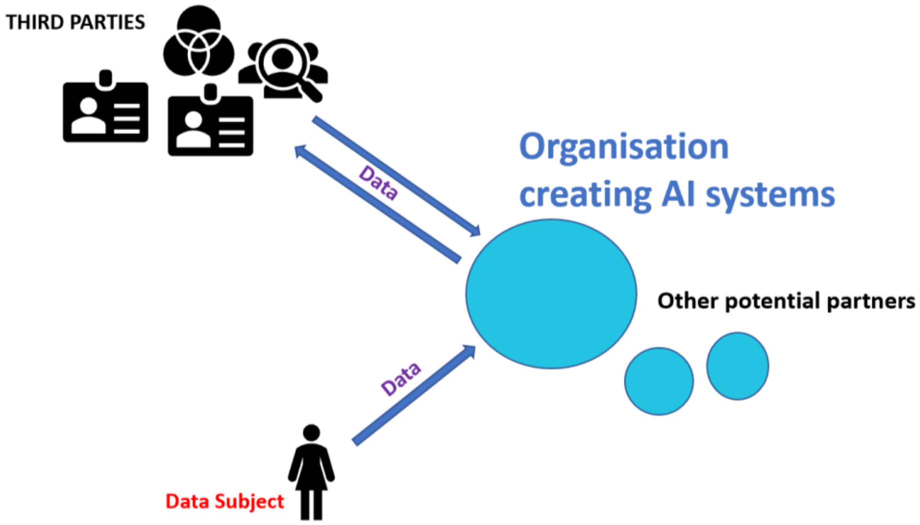


(Source: Griffiths et al., 2018)

A seeming lack of internal processes is generally counterbalanced by the dependency on external networks of third parties (Figure 8-3) whose practices often remain opaque or invisible.

Therefore, organisations should also assess how their data is processed by their external stakeholders. External relations demand meticulous checks for unethical and non-compliant practices which can affect compliance and reputation. Opacity and lack of transparency should not only concern algorithms.

Figure 8-3 Data exchange with external stakeholders



(Design: Chiara Addis)

b. Internal stakeholders

Fairness, accountability, and transparency of organisations are directly affected by what is done within teams. Internal teams 'practices need to be fair, accountable, and transparent. Internal accountability mechanisms should be assessed, and practices should be transparent and clearly understood, with no ambiguity around who does what, with which data, how and why. Such considerations and checks are especially important when organisations are creating ML systems using data from different internal data feeds, often assumed to be flawless following the checks expected by other teams.

Adequate internal mechanisms are required for data used as input data and for data created by the system (output data).

c. Special categories of internal stakeholders

Some areas, teams and roles bear more responsibilities around AI and DP, for example, Boards, Information Governance, DPO, IT and Digital teams. Their competences, composition and processes should consider the potential and risk of AI. They should have an understanding of AI and emerging technologies, be given effective power, and be involved in different moments of the innovation cycle. Their involvement in the pre- and post-project implementation phases (BAU) could help to reduce risks and increase compliance. Their expertise would be needed in three different phases (fig 8.1):

1. Prior to the development of projects — when the ideas for AI projects are generated, their expertise would contribute to increasing potential/mitigating risks of AI from the start.
2. During the development of the project — for example, at the beginning when this is assessed for approval (e.g., DPIA).
3. Post project/BAU — when checks on the new system ensure functionality and compliance (e.g., to satisfy Data Subject Rights).

Therefore, more involvement and more effective power are needed in the full innovation circle. Such power should not be limited to assessing ideas and vetoing projects but should encompass strategic and operational aspects.

AI is not only an IT and digital matter but a multidisciplinary effort. The expertise and competence existing in different parts of the organisation (e.g., Equality and Diversity) are a resource for AI projects, and they should be involved. As different subjects and teams can be used for different terminology and practices, their understanding and perceptions need to be assessed for a common understanding.

Users, unions, or other internal groups of staff are other important internal stakeholders with the power to influence projects. They should be informed and involved as much as possible.

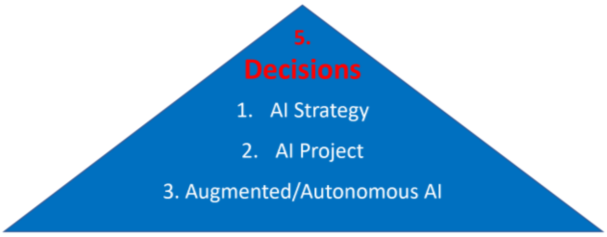
However, telling users what data will be processed without informing them about its level of granularity is not enough. They should be given all details, e.g., what, and how much data is used, its granularity, the exact purpose, the prediction and how this informs the final decision (see 8.5).

Finally, FAT and ethics do not belong to specific people or roles. Everybody should be involved and feel part of that milieu creating innovation. Staff who do not feel themselves to be included resist, and this impacts upon compliance, continuous improvement, and transformational change. Responsible management should be a non-performative/non-box-ticking collective exercise.

8.4.5 Decision-making in AI&DP management

In the following, the important elements around decision-making are presented. Organisations focusing on adopting responsible AI and DP management should consider how decisions are made. This research identifies 3 different key areas, AI strategy, AI projects, and Automated and Augmented AI (Figure 8-4).

Figure 8-4 Decision making key areas



(Design: Chiara Addis)

a. Organisational AI Strategy

While deciding the organisational AI strategy, careful consideration should be given to the motivation behind the acquisition of AI and its implications for DP. AI can be highly disruptive for organisations. Such analysis should focus on the readiness and compatibility of the organisation with AI, e.g., the availability and reliability of existing resources (people, data, IT), processes, and business models.

Adopting AI without a deep analysis of its ramifications can affect Data Subject rights and impact on the future sustainability of the organisation (e.g., economic losses or reputational damage).

b. AI Project

Organisations are called to make various decisions regarding their AI projects, for example, the decision to carry out an AI project in the first instance, to change it, or to stop it entirely. Below, some of these decisions are discussed:

1. Decisions around the aims and the impact of the project. Whereas the project may aim to increase efficiency or reduce costs, other possible consequences should be carefully evaluated as they could affect the organisation.

   For example, the project could impact on the workforce. Would staff be retrained or made redundant? Would other consequences, e.g., on societal values, be assessed?

Which criteria are deemed to indicate success and how much flexibility is expected during and after the implementation? Are DP roles involved, and how much? How will the impact on other internal areas be considered? Could that trigger an internal process towards vetoing or stopping the project? Would the resulting economic loss be acceptable? Are there some fallback plans?

These are just some of the options and dilemmas the management should consider in the early stages of an AI project.

2. Decisions around the type of AI, data, third parties. For example, deciding to use a third party to develop an ML Augmented system. This includes making decisions around responsibilities and accountability, and avoid situations where, for instance,

    - Developers do not know the specific context in which the system will be implemented.
    - Those implementing the system do not know how it works and avoid asking questions due to the perceived complexity of AI.
    - Both developers and those implement it thinks the other is responsible for its outcomes.

3. Decisions around control and safety measures (e.g., downgrading/stopping a project). For example, deciding if, when and how to perform an assessment, or when to retrain the algorithms, or the criteria for stopping automated systems in the case of unexpected outcomes or reputational damage.

    While changing or stopping a project indicate the high maturity and responsibility of organisations, such trade-offs come with high costs that many organisations may not be willing to pay. Such decisions require a high level of responsibility.

    Another crucial decision regards the repurpose of the system, a growing and risky practice. Later use of the system for a new purpose should not be possible without the assessment and authorisation of the persons in charge of DP.

c. Using AI to make decisions. Augmented AI and Automated AI models

Other decisions are linked to the specific AI model chosen by organisations for their AI decision-making processes, either AI to replace human decisions (Automated AI) or to support them (Augmented AI).

Automated AI systems make decisions without human intervention and require careful consideration. They are strictly regulated by the GDPR, prescribed only in some cases, and Data Subjects are given specific protection.

While Augmented AI systems are not regulated by the GDPR, they are often preferred as they are considered less risky (as seen in CS1). Even though the presence of humans can provide more control, the risks associated with this model are often underestimated.

Due to the different levels of protection, the decisions made by the two models should be clearly differentiated, and this can be challenging. Organisations need to demonstrate that the role of the human is significant and not a mere certification of the prediction. In such cases, decisions would be de facto automated decisions, requiring a higher level of protection. Conversely, if the processing by the Autonomous system is presented by an organisation as "necessary" for its business model, this could provide, in theory, the lawful reason for processing personal data without the consent of Data Subjects. However, the organisation would need to demonstrate the necessity of processing, something which could be difficult to prove with the ICO or with courts.

Therefore, organisations should plan very carefully how to use AI for their decision-making processes.

As Augmented AI poses specific risks for responsible management, some detailed suggestions will be extensively presented in the following section.

d. Responsible Augmented AI

The AI-Human interaction is at the core of Augmented AI. The "human in the loop" makes the decisions at the end of a long decision-making process encompassing two parts (Figure 8-5).

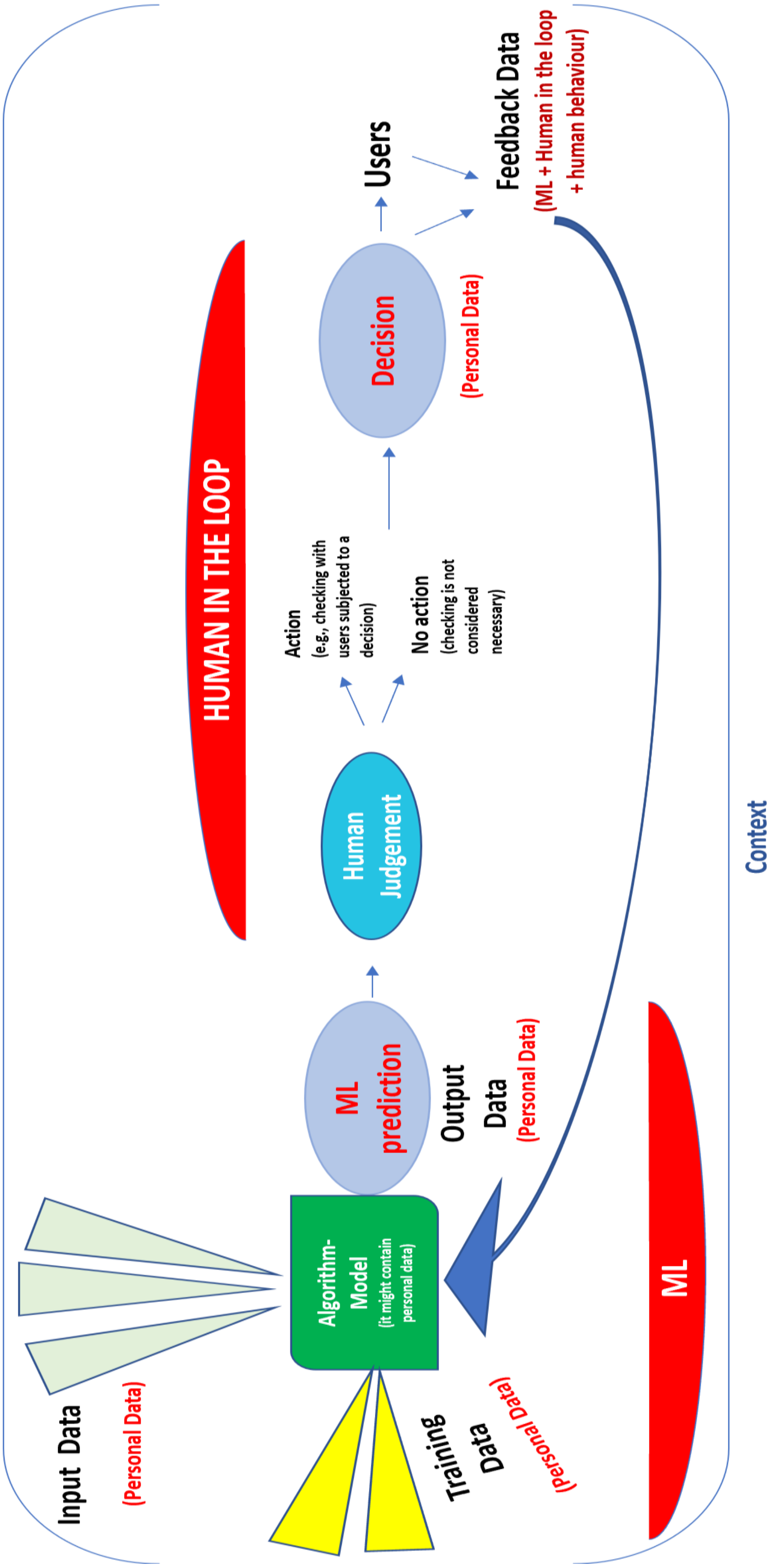
- In the first part, ML generates the information. The algorithm uses the input data to calculate the prediction (output data).
- In the second part, the human intervenes interpreting the prediction and making a judgement on top of that information. They may decide to integrate that information by acquiring new details. For example, a judge could choose to meet a potential re-offender flagged up as high risk by ML and make the decision on their release only after the encounter.

> Additionally, while such a decision is one impacting the Data Subject/user, it also has an effect, via feedback data, on the organisation. This new learning for the ML system is the result of information created during the various stages of the process – prediction, human interpretation, human decision, and its impact on the user in that context. Figure 8-5 illustrates the complexity of such a decision-making process, highlighting the personal data used and generated during the process (training, input, output, decision).

Therefore, Responsible AI&DP Management must consider all elements of this process capable of creating unfair outcomes. For example:

1. The data used to train the algorithm and the input data could be biased.
2. The model could be wrong. E.g., the weighting of the ML model could be incorrect, or the ML type (e.g., unsupervised learning) could be inadequate for that organisational need.
3. The judgement of the prediction could be incorrect. E.g., prediction read as a certain forecast, or a pattern between data read as causation (and not a correlation).
4. The judgements and decisions could be made without:
    - Considering the context of past events (e.g., economic recession), or the specific situations of a Data Subject (health issues) — specificity of the past.
    - Considering the context of current events. E.g., if a negative decision is made only considering the past behaviour of the Data Subject, without considering new and unexpected situations (e.g., the impact of Covid-19 pandemic on personal finances or student attendance) — the specificity of the present.
    - Looking for further information or a second opinion when doubts emerge (e.g., if not meeting the Data Subjects for further details).
    - Having the effective power to override the prediction, and merely confirming the prediction.

Figure 8-5 Augmented AI decision making and the human in the loop



(Design: Chiara Addis)

Therefore, decisions made with an Augmented AI system are the result of the combination of knowledge (and values) embedded in the data, and knowledge (and values) possessed by humans.

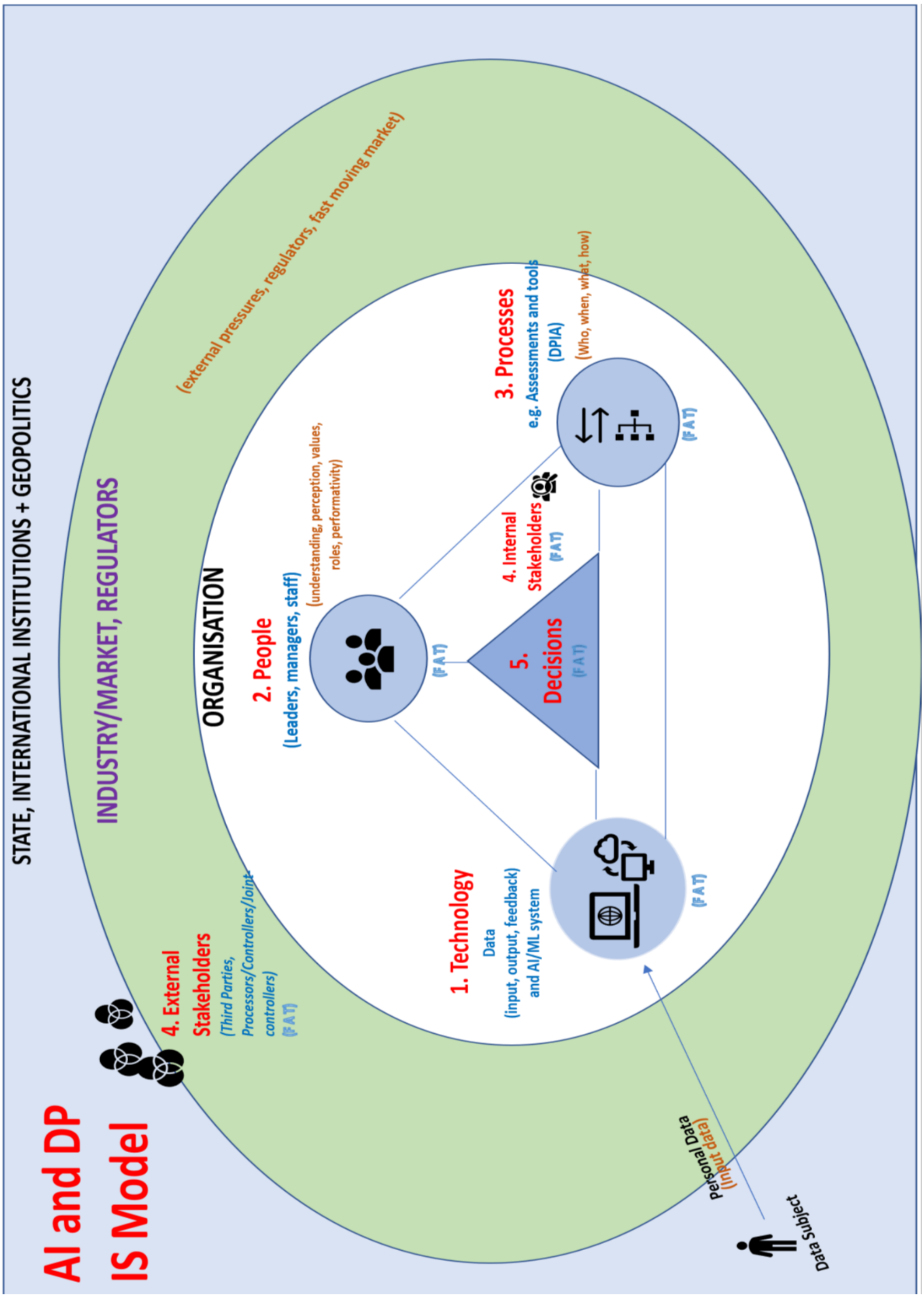
As algorithms are neither neutral nor objectively accurate, so can be the judgements and decisions of humans. Those aiming at eliminating unfairness and increasing ethics in the process should not only focus on the elimination of biases from the data, as this would not be sufficient. Factors like misinterpretation, lack of expertise and potential malign activities undertaken by humans cannot be solved by debiasing data.

Thus, organisations willing to use Augmented AI responsibly should consider such complexity and plan accordingly. For instance, transparency of both decision-process and final decisions should be sought and provided to final users, especially when different data sources and factors contribute to decisions organisations are accountable for. Additionally, the human in the loop needs competence and effective power. Thus, it would be advisable to increase training for staff, diversifying competences and backgrounds, and fostering lateral thinking and independent judgement.

8.5 *RAIDIS* management model - visual representation

In the following, a visual representation of the *RAIDIS* models is presented (Figure 8-6). This comprehends the five key elements (technology, people, processes, stakeholders, decision-making), the elements included in the external and internal organisational context, and how they are connected to each other. The figure helps to visualise the complexity of the ecosystem where innovation is created and the different elements that can impact on it, determining its responsible outcomes. Organisations interested in adopting Responsible AI&DP Management should take into consideration all elements included in such an ecosystem.

Figure 8-6 RAIDIS management model


AI and DP
IS Model
STATE, INTERNATIONAL INSTITUTIONS + GEOPOLITICS
INDUSTRY/MARKET, REGULATORS
(external pressures, regulators, fast moving market)
4. External Stakeholders
(Third Parties, Processors/Controllers/Joint controllers)
(FAT)
ORGANISATION
1. Technology
Data
(input, output, feedback)
and AI/ML system
(FAT)
2. People
(Leaders, managers, staff)
(understanding, perception, values, roles, performativity)
(FAT)
3. Processes
e.g. Assessments and tools
(DPIA)
(Who, when, what, how)
(FAT)
4. Internal Stakeholders
(FAT)
5. Decisions
(FAT)
Personal Data
(Input data)
Data Subject

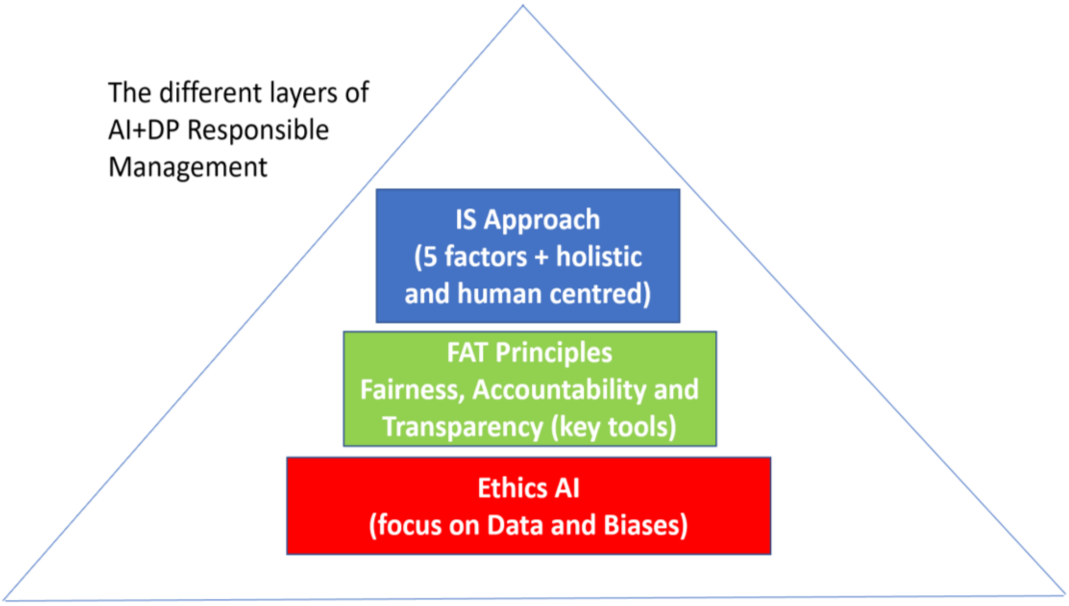
(Design: Chiara Addis)

8.6 Data Ethics, FAT, and Responsible AI&DP management

Organisations interested in the adoption of Responsible AI&DP Management should consider its diverse composition (fig 8.7) concerning ethics.

The current discourse on AI ethics is mainly Data Ethics, mostly focused on specific data and biases in data. Within such debate, the FAT principles are used to inform the choices made in the process, and can become precious instruments in the implementation and use of AI.

Figure 8-7 Complexity of responsible AI&DP management



(Design: Chiara Addis)

Therefore, the researcher considers Responsible AI&DP Management as an approach inclusive of three components - Data ethics, FAT, and the holistic IS framework – which can effectively inform the management of practical ethics.

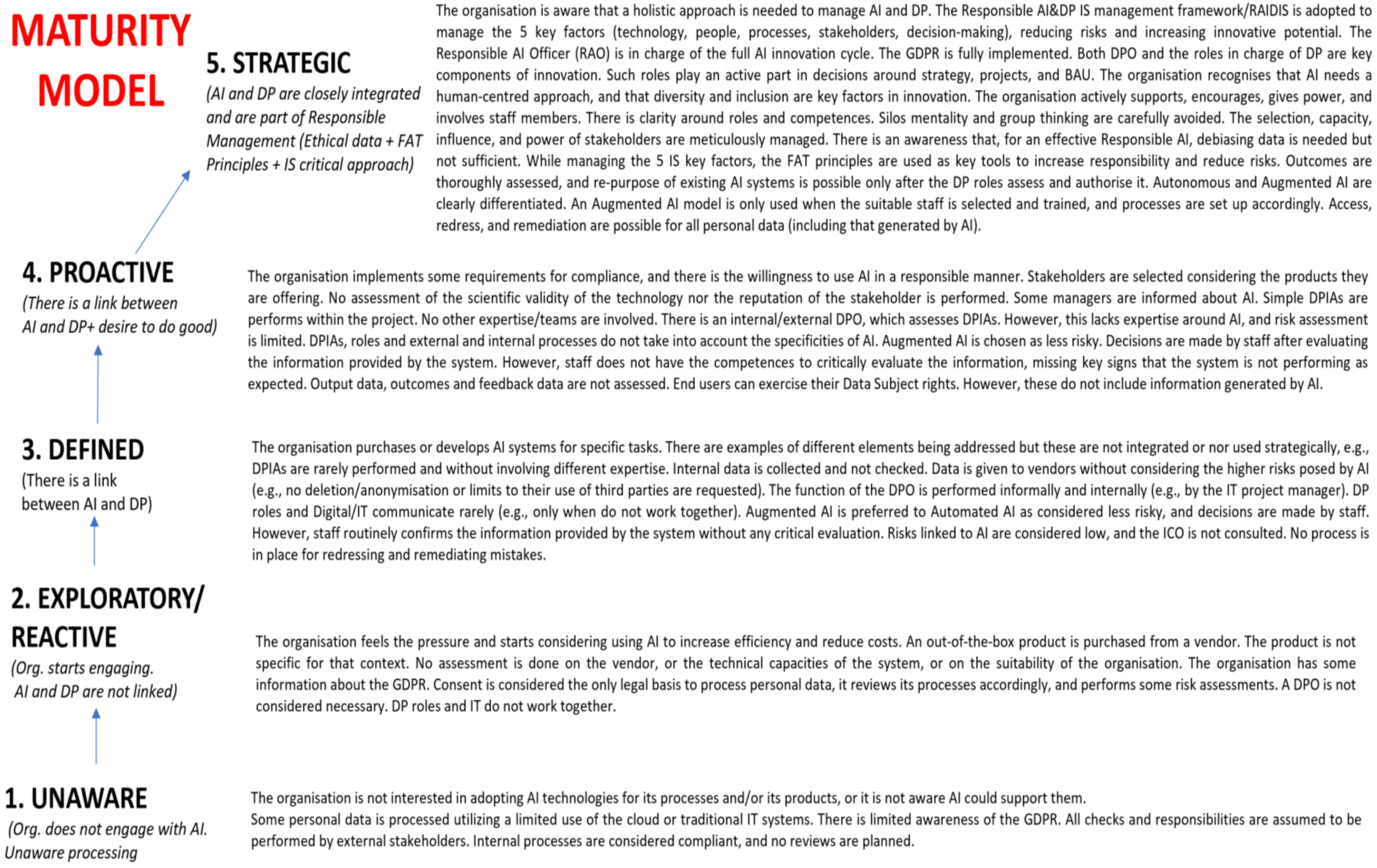
8.7 *RAIDIS* maturity model

This section presents a maturity model which illustrates how AI and DP can be included in the organisational strategy. The model draws upon the work carried out by Stahl et al. (2017) and Fenton, Fletcher, & Griffiths (2019). The first elaborated an RRI maturity model for practical implementation and action, while the second emphasised the importance of a holistic and system-focused approach, aiming at an aspirational future strategic level where maturity is understood as a continuous process.

The *RAIDIS* maturity model is created by identifying five stages of maturity and by linking them with some insights from the research. The result shows the evolution in the adoption and use of AI and DP, and some of the limits in the intermediate stages.

At the lowest level of maturity, the organisation has a limited interest in or knowledge of AI and the GDPR. Following external pressure from competitors or internal teams (e.g., IT), the organisation can react by buying the technology, without questioning its capability. Such a decision is often made without involving other areas. There is low awareness about the GDPR, only a few requirements are known (2). In further stages, the uses of AI become clearer, and different areas start working together, first for more specific tasks, then for more strategic aims. The GDPR and its requirements slowly become more important also for data management and strategy. The discourse around AI ethics slowly evolves towards a more responsible approach inclusive of FAT principles. The increased knowledge of AI and DP has a direct impact on the AI products used by the organisation. For instance, in the first stages of the maturity model, AI could be used to improve the recruitment process of an organisation, e.g., selecting candidates via video interviews. The system could include Emotion AI to classify the suitability according to personality traits, honesty, and reliability. In more mature stages, the use of such systems could be questioned, challenged, and dismissed for being potentially discriminatory and based on pseudoscience.

Figure 8-8 RAIDIS MM



MATURITY MODEL

5. STRATEGIC

(AI and DP are closely integrated and are part of Responsible Management (Ethical data + FAT Principles + IS critical approach)

The organisation is aware that a holistic approach is needed to manage AI and DP. The Responsible AI&DP IS management framework/RAIDIS is adopted to manage the 5 key factors (technology, people, processes, stakeholders, decision-making), reducing risks and increasing innovative potential. The Responsible AI Officer (RAO) is in charge of the full AI innovation cycle. The GDPR is fully implemented. Both DPO and the roles in charge of DP are key components of innovation. Such roles play an active part in decisions around strategy, projects, and BAU. The organisation recognises that AI needs a human-centred approach, and that diversity and inclusion are key factors in innovation. The organisation actively supports, encourages, gives power, and involves staff members. There is clarity around roles and competences. Silos mentality and group thinking are carefully avoided. The selection, capacity, influence, and power of stakeholders are meticulously managed. There is an awareness that, for an effective Responsible AI, debiasing data is needed but not sufficient. While managing the 5 IS key factors, the FAT principles are used as key tools to increase responsibility and reduce risks. Outcomes are thoroughly assessed, and re-purpose of existing AI systems is possible only after the DP roles assess and authorise it. Autonomous and Augmented AI are clearly differentiated. An Augmented AI model is only used when the suitable staff is selected and trained, and processes are set up accordingly. Access, redress, and remediation are possible for all personal data (including that generated by AI).

4. PROACTIVE

(There is a link between AI and DP+ desire to do good)

The organisation implements some requirements for compliance, and there is the willingness to use AI in a responsible manner. Stakeholders are selected considering the products they are offering. No assessment of the scientific validity of the technology nor the reputation of the stakeholder is performed. Some managers are informed about AI. Simple DPIAs are performs within the project. No other expertise/teams are involved. There is an internal/external DPO, which assesses DPIAs. However, this lacks expertise around AI, and risk assessment is limited. DPIAs, roles and external and internal processes do not take into account the specificities of AI. Augmented AI is chosen as less risky. Decisions are made by staff after evaluating the information provided by the system. However, staff does not have the competences to critically evaluate the information, missing key signs that the system is not performing as expected. Output data, outcomes and feedback data are not assessed. End users can exercise their Data Subject rights. However, these do not include information generated by AI.

3. DEFINED

(There is a link between AI and DP)

The organisation purchases or develops AI systems for specific tasks. There are examples of different elements being addressed but these are not integrated or nor used strategically, e.g., DPIAs are rarely performed and without involving different expertise. Internal data is collected and not checked. Data is given to vendors without considering the higher risks posed by AI (e.g., no deletion/anonymisation or limits to their use of third parties are requested). The function of the DPO is performed informally and internally (e.g., by the IT project manager). DP roles and Digital/IT communicate rarely (e.g., only when do not work together). Augmented AI is preferred to Automated AI as considered less risky, and decisions are made by staff. However, staff routinely confirms the information provided by the system without any critical evaluation. Risks linked to AI are considered low, and the ICO is not consulted. No process is in place for redressing and remediating mistakes.

2. EXPLORATORY/ REACTIVE

(Org. starts engaging. AI and DP are not linked)

The organisation feels the pressure and starts considering using AI to increase efficiency and reduce costs. An out-of-the-box product is purchased from a vendor. The product is not specific for that context. No assessment is done on the vendor, or the technical capacities of the system, or on the suitability of the organisation. The organisation has some information about the GDPR. Consent is considered the only legal basis to process personal data, it reviews its processes accordingly, and performs some risk assessments. A DPO is not considered necessary. DP roles and IT do not work together.

1. UNAWARE

(Org. does not engage with AI. Unaware processing

The organisation is not interested in adopting AI technologies for its processes and/or its products, or it is not aware AI could support them.
Some personal data is processed utilizing a limited use of the cloud or traditional IT systems. There is limited awareness of the GDPR. All checks and responsibilities are assumed to be performed by external stakeholders. Internal processes are considered compliant, and no reviews are planned.

(Design: Chiara Addis)

8.8 How the *RAIDIS* management and maturity models can be used

Both *RAIDIS* management and maturity models can be used for strategic, risk, and stakeholder management. They can be adopted internally when:

- Defining AI strategies, in order to understand the resources available, and evaluate the impact of potential projects on the organisation.
- Defining, implementing, and evaluating potential projects, and selecting third parties/vendors.
- Monitoring, controlling, and evaluating the use of AI in BAU.

Both models can also serve as strong supporting tools while dealing with

- Regulators. E.g., the ICO. The models can show the specific and rigorous approach followed while implementing and using AI. This can strengthen the position of organisations by demonstrating their accountability.
- Funding bodies, investors and shareholders. As the demand for corporate environmental, social, and governance (ESG) practices is growing, having strong models capable of shaping responsible practices and clearly demonstrating their evolution can increase interest, trust and investments in the organisation.

8.9 Synthesis: *RAIDIS* management and maturity models

This section presents a synthesis of the key insights on Responsible AI&DP Management.

This synthesis can provide managers and leaders with an ideal type of organisation already using the framework. In this short version, the key insights of the model are presented as success factors. The following stages provided an idea of the gradual adoption of the success factors by an organisation.

- Stage 5: 81-100 % of success factors are present
- Stage 4: 61-80% of success factors are present
- Stage 3: 41-60% of success factors are present
- Stage 2: 21-40% of success factors are present

- Stage 1: 0-20% of success factors are present

This aligns to the stages of the maturity model. A simpler version of the maturity model is presented at the end of the section.

a. Technology

- All data — training, input, output, feedback data — is checked for biases.
- Organisational practices are also checked for potentially biased behaviours, ethical issues, and for any risk they can pose to responsible management.
- Both algorithm auditing and systems checks are regularly performed.
- There is awareness that AI can aggregate non-personal data and identify Data Subjects. Output data is assessed considering such risk.
- Biometrics are managed carefully due to the higher risk for Data Subjects, and the enhanced DP requirements. Both the acquisition and use of biometrics are carefully assessed.
- Management is aware that many AI products on the market lack scientific bases. Leaders know that many criteria used to classify people from physical characteristics (e.g., facial expressions, skin tone) lack scientific validity, are context-specific, and can be used for discriminatory practices. Strong checks are put in place to verify AI products.
- The scarcity of available high quality and varied data is not counterbalanced by an increased quantity of collected data. Quality and variety of data are carefully sought.
- ML is used for identifying past correlations in data and for making predictions (future guesses) when conditions are similar. ML is not used when past data is not clear, scarce, or new conditions arise. In such situations, the predictions performed by humans outperform those of algorithms.
- The organisation provides training and education at a different level in order to increase the awareness and the understanding of AI and DP.
- There is an awareness that data cannot provide all information regarding the complex experience of Data Subjects. These are consulted as much as possible.

- There is awareness that human behaviour is the result of many variables and that its prediction is difficult or impossible.
- There is an awareness that past data used to predict future behaviour is also the result of all factors present in that context, including unfair practices.
- The impact on different users is assessed, and remedies and redress are always offered to Data Subjects.
- To increase the accuracy and fairness of predictions, new personal data from contemporary events is integrated with data from past events.
- Organisations know transparency is an important GDPR requirement for their data and practices. They require the same transparency from their partners/third parties.
- Organisations require as much information as possible from vendors and developers, and carefully assess it.
- Organisations using Deep Learning are aware of its characteristics. They know that the decision-making process is not transparent and that they will be held accountable for such decisions.
- The contexts where AI is created and where it is deployed are clear. Their characteristics, the potential discrepancies and the related impact are assessed.
- Security includes cybersecurity, internal and external threats, and the understanding that data can be manipulated and not only stolen.

b. People

- A *Responsible AI Officer (RAO)* is appointed. They possess a strong understanding of AI, DP, and responsible innovation, and have the full responsibility for the complete AI innovation cycle (from strategy to BAU). This increases the effectiveness of the holistic framework.
- Competences, responsibilities, and boundaries between roles are very clear.
- The environment is multidisciplinary and diverse, and there is a balance between technical and non-technical roles.

- High expertise and participation from different areas and disciplines are sought to reduce biases.
- Everybody receives some training in AI and DP tailored to context and role. Their understanding is verified.
- People are given spaces for exchanging knowledge and expressing opinions and suggestions around responsible innovation.
- People are part of the innovation process. They feel part of that, and they feel listened to.
- Top-down approaches are avoided.

c. Processes

- All different moments within the innovation cycle (strategy, project, BAU, and impact on organisation) are considered and assessed using different tools.
- The assessments (e.g., DPIAs, equality and algorithm assessments) are considered key innovation tools. They are not perceived as box-ticking exercises, but important moments where different expertise comes together for evaluating risks and exchanging ideas and contribute to innovation. They are not only performed within projects, but at other moments as well (e.g., BAU).
- New and existing processes and related documents are adapted considering the acquisition and the use of AI. Forms for performing assessments include specific requirements for AI, and different versions of documents exist for different areas (e.g., more and less technical).
- Long term sustainability of all human and material resources is considered and planned carefully.

d. Stakeholders

- There is an awareness that collaboration with third parties can be very beneficial but also potentially very risky. Their competences, knowledge, and technical capabilities are always assessed carefully. Specific documentation around their services, reputation and information on past products is requested and assessed for technical expertise, DP, responsible practices, and reputation.

- Contracts with third parties are very specific around competences, DP, and multiparty liabilities. When third parties use other entities/third parties to process data provided by the organisation, the details of the processing are clearly included in the contract.

- Technical and DP practices of internal stakeholders are assessed. Specific attention is given to the areas where wrong assumptions or perceptions can occur more frequently (e.g., internal accountabilities and quality of data).

- Roles, competences, and knowledge are updated taking AI into account.

- The power and expertise of the people more knowledgeable around AI and DP are not limited to the pre-approval stage of the projects. They also influence strategies, project work and BAU.

e. Decision-making

- There is clarity on the purpose of the innovation. Re-purposes are carefully assessed for DP compliance by the DP roles. Re-purposes are not possible without their authorisation.

- There is an awareness that the decisions made around AI and DP by different people and third parties can impact on the responsible practices of the organisation and on the rights and freedom of Data Subjects.

- Decision making is clearly defined and planned with regard to the three domains of strategy, projects and use of AI.

    1. Strategy. Organisations thinking of adopting AI systems evaluate the idea carefully. They take the time to assess their organisational and technical resources, the compatibility of AI with their aims and values and its impact on Data Subjects. After deciding on the acquisition, organisations invest in upgrading their technical and organisational resources. Increasing awareness and knowledge around AI and GDPR are considered necessary for responsible management. Effective participation and power of the roles involved in the innovation process are sought and supported.

    2. Projects. Organisations implementing AI projects take the time for identifying and choosing:

- The best AI technology for their specific needs (e.g., Deep Learning, ML, etc)
- The data to be used
- The most suitable model (e.g., augmented system which supports human decisions vs automated decision making).

Assessments (e.g., DPIA) are performed via different tools and involve various roles. Safety criteria for downgrading or stopping the projects are planned, for instance, in the case of unethical and pseudoscientific predictions or personal data breaches.

3. Use of AI. Organisations know that decisions made via AI can have different levels of complexity and transparency.

   Automated AI decision making systems processing personal data can be used when prescribed by the GDPR and are carefully monitored and controlled. Provisions are made to satisfy the GDPR enhanced protections (e.g., right to obtain human intervention).

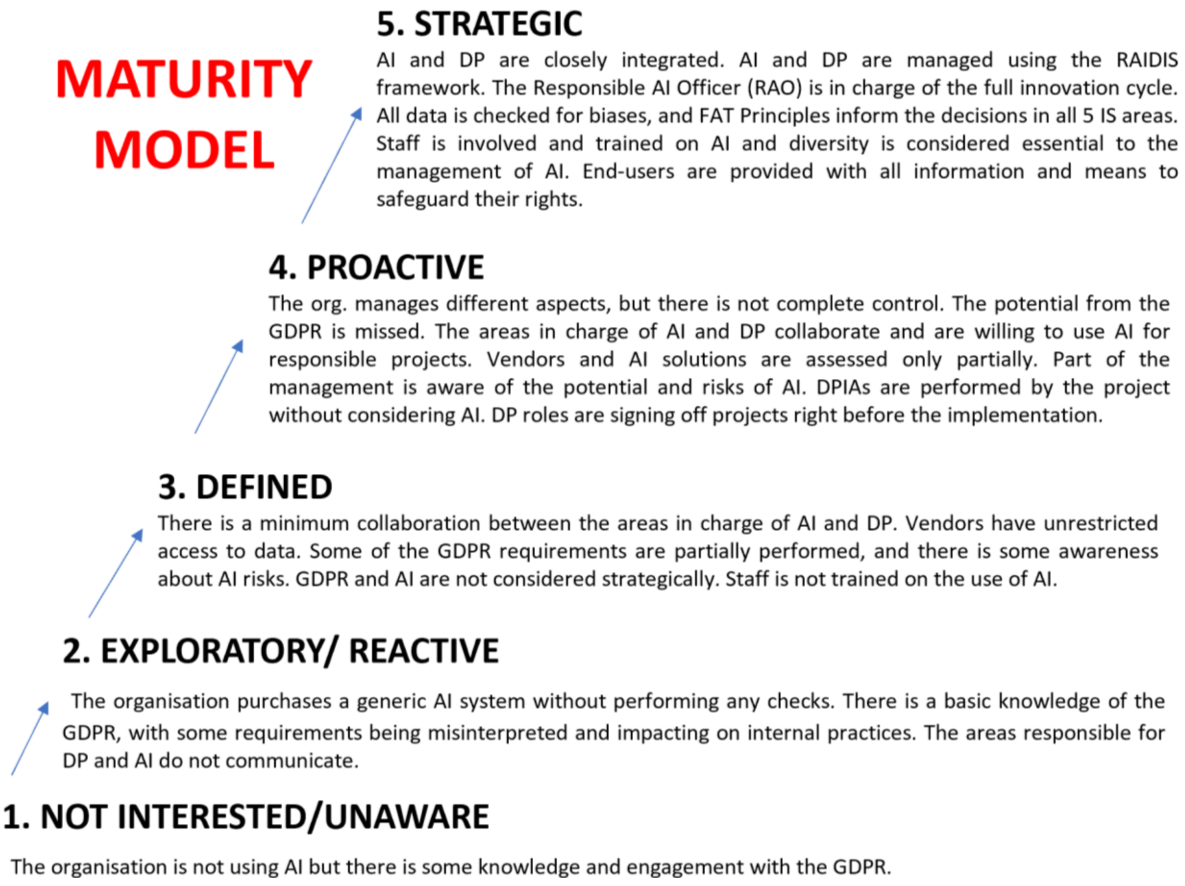
   Organisations using Augmented AI know that the final decision is the result of choices made during the whole decision-making process. Such choices can produce unfair outcomes, and they are carefully overseen. The capacities and the power of those making decisions (on top of ML predictions) are accurately assessed. There is an awareness that humans can miss biases in data and processes, misread predictions, and make wrong judgements and decisions. Humans using AI are fully trained, continuously supported and are given the effective power to make decisions, and to override the ML predictions whenever necessary.

   Due to their different legal requirements, the decisions made by Automated AI and those made with Augmented AI are clearly differentiated.

f. *RAIDIS* MM (reduced version)

A reduced version of the maturity model is presented in the figure below.

Figure 8.9 RAIDIS MM - reduced version



MATURITY MODEL

5. STRATEGIC

AI and DP are closely integrated. AI and DP are managed using the RAIDIS framework. The Responsible AI Officer (RAO) is in charge of the full innovation cycle. All data is checked for biases, and FAT Principles inform the decisions in all 5 IS areas. Staff is involved and trained on AI and diversity is considered essential to the management of AI. End-users are provided with all information and means to safeguard their rights.

4. PROACTIVE

The org. manages different aspects, but there is not complete control. The potential from the GDPR is missed. The areas in charge of AI and DP collaborate and are willing to use AI for responsible projects. Vendors and AI solutions are assessed only partially. Part of the management is aware of the potential and risks of AI. DPIAs are performed by the project without considering AI. DP roles are signing off projects right before the implementation.

3. DEFINED

There is a minimum collaboration between the areas in charge of AI and DP. Vendors have unrestricted access to data. Some of the GDPR requirements are partially performed, and there is some awareness about AI risks. GDPR and AI are not considered strategically. Staff is not trained on the use of AI.

2. EXPLORATORY/ REACTIVE

The organisation purchases a generic AI system without performing any checks. There is a basic knowledge of the GDPR, with some requirements being misinterpreted and impacting on internal practices. The areas responsible for DP and AI do not communicate.

1. NOT INTERESTED/UNAWARE

The organisation is not using AI but there is some knowledge and engagement with the GDPR.

(Design: Chiara Addis)

This synthesis of an ideal type of organisation already using the *RAIDIS* model permits to better visualise the various elements of the approach.

8.10 Conclusion

This chapter presented a proposal for responsible AI and DP management.

The chapter started by introducing an enhanced Information System (IS) framework used as an analytical tool. This includes five factors (technology, people, processes, stakeholders, decision-making) considered as key components of the *RAIDIS* management model, a holistic approach for a responsible implementation and compliance of DP/GDPR and AI projects.

The chapter then presented *RAIDIS* MM, the maturity model that illustrates the various stages towards a responsible organisational strategy. That was then followed by a synthesis of the model. By taking into consideration the characteristics of the context, and by providing specific suggestions of what needs to be done in each of the five factors, the *RAIDIS* model complements existing general guidance on AI ethics.

The result is a strong management approach capable of reducing risks while increasing compliance, innovation, and sustainability.

CHAPTER 9: CONCLUSION

9.1 Introduction

This final chapter presents the conclusion of the research, in accordance with the research question and research aims. After focusing on the discussion of the contributions of the research and an evaluation of its limitations, the chapter provides some personal reflections and ends with suggestions for future research.

9.2 Research summary

The research presented in this thesis has been concerned with ethical issues around the use of AI and DP. The growing diffusion of AI has sparked a debate on its uses and potential problems. The GDPR imposed new requirements for organisations processing personal data, some of which aimed at regulating the processing performed by AI systems. This research explored issues around the responsible management of AI systems by focusing on the interplay between AI, DP, and FAT principles.

The review of the literature discussed the most relevant work carried out on AI, DP, FAT, and innovation management. After presenting the development of AI and the role of ML, the review addressed the evolution of DP as a human right in Europe, the GDPR, Data Protection Act 2018, the UK GDPR, and the three FAT principles (the GDPR came into effect during the time of the PhD). It then introduced a reflection of the work on the management of digital innovation and the specificities of AI management.

The review highlighted same gaps in the literature, with some key elements missing from the debate. While the general focus in the literature is on the responsibility of AI developers, other stakeholders have not been considered as active elements in shaping AI and influencing its outcomes. Ethical issues related to the implementation and use of technology within an organisation and the roles of people, their power, and agency in their specific contexts have largely been missing from the discussions. Furthermore, fairness, accountability, and transparency clearly emerged as complex concepts not simply concerning data, but also the complicated system in which people, technologies and processes interact.

The methodology comprised a multi-method approach, deploying a number of interviews with experts, and two case studies of organisations implementing responsible AI projects, consisting of interviews with key subjects (20), document analysis and observations.

The experts provided extremely valuable information from within different sectors and insights related to their current assignments, while the case studies offered multiple data sources and in-depth insights into those settings. The combination of the two elements enabled a wide-ranging understanding of organisational practices and how various subjects (leaders, senior managers, DP, and ML experts) understood and perceived AI, GDPR and FAT.

Drawing on responsible research and innovation (RRI) and Andrew Feenberg's critical theory of technology, the research elaborated its own distinctive approach, i.e., critical AI&DP/*CRAIDA* Management. This permitted the exploration of the praxes of GDPR and AI management within UK organisations, including the processes and experiences, power positions and roles of various subjects and stakeholders inside and outside the organisations.

9.3 Evaluation of the research

The aims and the objectives of the research were generally met. Building upon a literature review of the most relevant and innovative sources, and by unveiling and exploring different aspects of management practices through original research and case studies, the research was able to achieve all its aims and objectives (Table 9.1).

Table 9.1 Aims and objectives of the research

AIMS			
1.Understand the relationship between AI and DP and how they can inform each other in the context of legislation and digital innovation	2. Examine the extent to which individuals who are introducing /using* AI and DP roles understand AI, DP, and FAT principles	3. Understand the impact of DP on organisations that are introducing/using AI, and vice versa	4. Produce guidance for organisations to support the application of FAT principles in their AI&DP Management

OBJECTIVES		
1. To identify how DP legislation protects personal data when processed by AI.	2. To investigate the level of understanding amongst AI adopters and users and DP roles, specifically: a. their knowledge, interpretations and perceptions of AI, DP, and FAT principles. b. whether the FAT principles are taken into consideration when AI systems are chosen, implemented, and used. c. how they use personal data, how they plan to use it, and the current and potential future impact of this on their organisations. (*internal users and not end-users)	3. To develop a critical theoretical framework that permits the unveiling of the innovation environment, and to produce a model on FAT principles aimed at supporting organisations in their AI&DP Management.

In the sections below, I will explain in closer detail how exactly the aims and objectives were achieved and what knowledge has been gained through the study. This will be done by discussing four different areas in aims and objectives:

a. Aim 1 and objective 1. The focus is on the relationship between AI and DP, and on the protection offered to personal data by the current legislation.

b. Aim 2, objective 2. The focus is on the experiences of subjects, their knowledge, understanding and perceptions.

c. Aim 3, objective 2. The focus is on the current and future impact of AI and GDPR on organisations.

d. Aim 4, objective 3. The focus is on producing guidance for organisations and providing practical tools.

a. The AI-DP relationship and the protection offered by the current legislation (aim 1, objective 1)

The research highlighted important aspects in the relationship between AI and DP, including how they can inform each other (aim 1), and how existing norms protect personal data (objective 1).

The research revealed how the rapid diffusion of AI is the result of the role played by multiple factors. The extremely rapid pace of the development of AI technologies, the high pressure coming from the market, and the increasingly easy and cheap modalities in acquiring the technology are all key factors in the AI hype of the last few years.

Conversely, the evolution of the DP legislation appears to be rather different. The legislative outcomes of the GDPR have been rather remarkable, constituting a turning point in the European DP history, whose centrality of the DP as a human right is its distinct feature. The GDPR has been not less important internationally, considering its extra-territorial reach and its influence on other DP regimes. And yet, its long legislative process is extremely slow considering the pace of innovation technologies. The Regulation was approved after four years of mediations and became enforceable after two years, a very long time in comparison to the pace of AI.

Furthermore, legislation needs enforcement. The low compliance in UK organisations is also strictly connected to the low enforcement measures taken by the ICO. What has happened in the UK after 2018 shows that the lack of robust enforcement measures can undermine strong legislation. Some critics of the GDPR question its success as an effective piece of legislation. And yet, such critiques do not appear to consider the fast pace of digital innovation nor the ICO's poor enforcement rate. Additionally, the politicisation of the GDPR through the lens of Brexit does not help to increase the culture of compliance around the Regulation.

Therefore, the fast development and use of AI on the one side have been accompanied by a long legislative process, slow implementation, and low enforcement on the other GDPR-related side.

The protection of personal data offered by the GDPR is extensive. By expanding the definition of personal data, the Regulation offered protection to various types of data and made provisions for the processing carried out by AI. Further clarifications were later provided by the DPA 2018.

Some further issues have emerged following the evolution and application of AI. Concerns around biometrics, emotions and mental data, data merging, and group inferences are growing. While biometrics are regulated, the protection of emotions is unclear, and the implication for AI systems that can allegedly track emotions is troubling. Furthermore, the devices tracking and analysing mental data of various subjects (such as employees) are increasing. While mental data is included in the GDPR definition, a separated and stronger protection is urgently needed with regards to privacy and DP.

Data merging is another issue. The growing capacity to infer information on Data Subjects and groups is concerning, especially considering the use of information collected from social media. CS2 is a small example of the power that online information can have in shaping reputations, and the recent case of Clearview (as discussed in 2.2.5) is a demonstration that existing DP rules, such as the ones around the public domain, cannot always provide clear protection, especially when such applications are already in use by police authorities in different countries.

Additionally, such cases show how oblivious Data Subjects can be while others are making decisions using that data. That increases the need for more recognition and power given to advocacy groups.

While this option is included in the GDPR, the way the UK detailed such representations excluded the representation of the interest without the Data Subject's mandate (Art. 80.2 UK GDPR). This de facto reduces the protection by denying independent organisations the ability to act on behalf of individuals and to realise collective redress/class actions.

Another critical issue revolves around the usage of AI to make decisions in automated and augmented systems and predictions. While automated decision-making is regulated by Art. 22 GDPR and by the details provided by the Data Protection Act 2018, giving Data Subjects specific rights, the decisions made with augmented systems lack such protection. This is critical, especially as this model is being preferred by many organisations as a less risky option. Data Subjects should be recognised similar protection to the one prescribed by Art. 22 for automated processing and know how the decisions affecting their rights were made by organisations. For example, they should know how the final decision was made, how influential the prediction was, and how much weight the human judgement had in making such a decision. This problem is closely linked to one of the main findings of this research: the common overestimation of the capacity of the Human in the Loop in the interaction with

AI. While human decision-making was considered the safer option, no provisions for training and support to humans making decisions were mentioned or considered by participants. The factors operating in that context and the experience and agency of the humans were not taken into consideration. Such issues highlight the need to expand these protections by also considering the operationalisation of AI and DP.

The analysis of the praxis highlighted similar issues, less dependent on the text of the GDPR. For example, what was considered to be personal data in different contexts, or how the GDPR was misread by some people, revealed the need to also focus on the understanding of the law during its implementation, i.e., on processes, competences, and the power given to DP roles.

Self-regulation would not be enough to regulate AI. The GDPR is a starting point, and it needs to be followed by legislation created via more flexible and faster processes, which considers what happens inside organisations. But strong compliance also needs robust enforcement.

b. Experiences of subjects (aim 2, objective 2)

The research examined the experience of those adopting and implementing AI, and their understanding and perceptions of AI, DP, and FAT principles (aim 2, objective 2).

The research highlighted how crucial the experience of individuals inside organisations was in shaping the responsible innovation created in such contexts. It revealed how the product of the innovation was the result of a combination of various factors. While the importance of technology was central, the role played by people in different moments of the innovation cycle was often underestimated. The research highlighted how knowledge, personal experience, roles, and performativity all play a part in the capacity of individuals to shape responsible innovation. It also unveiled a level of complexity and diversity in the experience of various participants, which was not expected. This demands a stronger focus on the role played by people in making sense, implementing, and using AI and personal data in their daily activities in future research.

Furthermore, the research showed how the capacity of people to understand or to perceive AI, DP and FAT principles should not be taken for granted. The capabilities of AI and humans can be overestimated within organisations, and project decisions can be more dependent on assumptions and perceptions than on scientific and regulatory factors. Besides, the definition of a responsible project aim can result in a false sense of reassurance regarding the assumed responsibility of the full management journey. Furthermore, the understanding of the FAT principles varied considerably amongst participants. What is fair for some, is not fair for others. While some processes or products were considered transparent by some participants, they could have been considered quite differently by the users.

The understanding of the principle of accountability also highlighted a crucial connection. A low sense of agency and self-reflection can impact on accountability. For example, people and organisations were able to see more easily the danger associated with practices performed by others and by other companies than the potential risks and dangers caused by their own actions. This "otherness of irresponsible practices" is significant and should be considered very carefully while managing stakeholders. Lack of awareness around power and power dynamics emerged in various moments. Such situations should be wisely considered by those willing to adopt responsible AI practices.

Additionally, the capacity to read, perceive and act responsibly needs to be sustained by organisations. This cannot be done only via top-down approaches. People need to trust the technology in order to use it, but they also need to trust the organisation and feel part of the innovation created in that context.

c. The current and future impact of AI and GDPR on organisations (aim 3, objective 2)

The research explored the current and potential future impact of AI and GDPR on organisations (aim 3, objective 2).

GDPR compliance was generally low, and very much dependent on the sector and the size of the organisation, and awareness of the people inside the organisation. A higher level of compliance was recorded in regulated sectors. However, the findings revealed different ways to be compliant. How power is enacted can be very different from how it is prescribed in the GDPR. Performing DPIAs, having DPOs or updating privacy policies does not necessarily imply compliance. Often the issues lie not on what is done but how. Performing assessments without considering all potential risks and only as a box-ticking exercise is not enough. Similarly, having a DPO which is not consulted, or consulted too late, does not provide the level of protection prescribed and hoped for by the Regulation. Moreover, the pace of change within socio-economic factors can make training data rapidly outdated, an external factor only mentioned by one participant. Furthermore, in some cases, compliance was considered difficult and highly impacting. But the complications were often ascribed to the wrong causes or a misinterpretation of the requirements. For example, although being crucial for the project described in CS2, the GDPR and its requirements were considered to be top-down obligations that were not always easy to satisfy. But then, some of the identified difficulties did not necessarily originate from the Regulation. Conversely, the GDPR did not appear as an issue within the project documented in CS1. It had already been implemented and some requirements were already performed prior to 2018. However, some implications were more concerning for well-informed leaders than for DP roles and project members. Therefore, the assertion that an organisation is compliant should always be assessed carefully.

The impact AI is having on organisations varies, and again this is not only dependent on the business model, maturity, and technology already in place, but also on the knowledge and awareness of insiders. In CS1 AI appears as an element to be added to existing processes in order to speed up decisions, an add-on for a well-structured organisation which wanted to keep control of the system and its use. In CS2 AI is seen as a key element of its business model based on the sharing of information between subjects and different entities, an external intelligence used ad hoc whose control is completely outside the remit of CS2.

Finally, a widespread lack of codified rules on new AI classifications and decisions made on top of ML predictions emerged. In addition, AI and DP were generally managed via traditional understandings of personal data and IT, and not via distinctive management adequate to their specificities.

The future impact is likely to be consistent. Education is moving towards personalised education (as seen in 5.2). While the project documented in CS1 has gradually reduced the space for ML predictions due to concerns about fairness and lack of control, discussions on the use of AI for personalised education were ongoing. There was not a plan, but discussions around the future of a personalised education built around the present and predicted needs of students were already there. This will raise serious questions about the role of external stakeholders and FAT principles in shaping future students as key citizens and political actors. Similarly, the types of services provided by the project analysed in CS2 will likely increase, fuelled by services on demand provided by ecosystems of entities whose business practices are completely oblivious and hardly accountable.

Such situations will demand different tools and approaches to legislation and management practices in order to protect the rights and the freedoms of individuals when AI is involved.

d. Guidance (aim 4, objectives 3)

The research also aimed at producing specific guidance on FAT and AI management (aim 4, objective 3). The research has produced two documents: Guidance in the form of a responsible IS AI&DP/*RAIDIS* model which presents a detailed picture of all key factors and their connections shaping responsible AI management. A detailed and long version and a short one were prepared, *RAIDIS MM,* a maturity model which presents the journey of an organisation towards responsible management, outlining a holistic approach to management that considers different elements which could be capable of creating unfair outcomes.

9.4 Contribution of the research

The research provides a significant original contribution to theory, knowledge, and practice

a. Contribution to theory - *CRAIDA* management framework

After addressing a gap in the existing theory, the research elaborated critical AI&DP/*CRAIDA* management framework*,* which addresses the management of the technology in practice. By drawing on RRI and Andrew Feenberg's CTT, the framework focuses the attention on the context, the experience of individuals in organisations, societal and personal values, the role of stakeholders, and power, risks, and processes.

Further factors contribute to make *CRAIDA* also a distinctive management framework. This is for the following reasons:

- The object of the exploration is not the research environment, but the organisation operationalising AI.
- The focus is not the development of AI, but its implementation and use.
- The main actors are all those shaping AI (not only developers).
- The external environment is also considered in influencing internal praxis (not only the internal one).

Therefore, *CRAIDA* provides a holistic and innovative human-centred framework theory that is capable of unveiling the hidden layers of the explored innovative milieu and highlighting how organisational practices and technology were shaped by the experience of the subjects.

b. Contribution to knowledge - *RAIDIS* model and responsible AI management

The research also created the responsible IS AI&DP/*RAIDIS* model, an instrument for organisations adopting AI responsible management. Drawing upon Information System Management, the model was created considering the socio-technical system where AI is used and where humans make decisions. The model provides a strong mechanism for governing complex innovative environments. By clearly illustrating five key factors (technology + data, people, processes, stakeholders, decision-making) and how their interrelations impact compliance and responsible innovation, it provides a practical and detailed instrument for identifying exact risks and potential within specific areas.

Furthermore, the research identifies the limits in the current discourse around AI ethics and makes a crucial connection between AI ethics, FAT principles, and responsible AI. AI ethics is mainly focused on data (input and training data) and biases. Contexts, output data, and the role of those implementing and using AI are largely unexplored. The research used the FAT principles to read and understand the experience of people, and to shape the practices around the five IS factors (technology, people, processes, stakeholders, and decision-making).

Doing all this, *RAIDIS* is established as a multi-layered approach that encompasses data ethics and FAT principles and provides organisations with a solid framework for holistic management.

c. Contribution to practice - *RAIDIS* Maturity Model and *Responsible Augmented AI*

The research created a maturity model for organisations adopting AI responsible management, and it further unveiled the complexity of decision making in Augmented AI models.

1. *RAIDIS MM*. The model presents the journey of an organisation towards responsible management and clearly illustrates how AI and DP can be included in organisational strategies. By illustrating practical implementation and actions to be taken in the five identified stages of maturity, it provides a holistic and system-focused approach aiming at a strategic maturity level where AI and DP are the responsibility of the management.
2. *Responsible Augmented AI*. Another contribution is provided around decision-making processes and the role of human judgement in Augmented AI systems. By clearly considering the different stages of the decision-making process (prediction,

judgement, action, and decision), the research identifies some specific AI and DP risks that could occur in each stage. This permits the better identification of exactly what is required in each moment to better manage risks and increase responsible management. Furthermore, it clearly identifies the two parts of the Human-AI Interaction; this combination of knowledge encompasses completely new dynamics. Such findings around the specificities of Augmented AI provide more clarity around the management of an AI model which is often chosen by organisations as it is perceived to be safer, but whose intrinsic risks can be underestimated.

9.5 Limitations of the research

The research has also some limitations, primarily due to the research and legislative contexts, the degree of maturity of the technology and of GDPR compliance. Additionally, despite careful planning, some limitations around the methodology emerged during the research journey.

a. Research and legislative context

This research has been done within the context of the European DP tradition, the GDPR, and the requirements the Regulation places on UK organisations. This focus creates some unavoidable limitations. The research did not explore the application of the Regulation by other EU Member States, nor the experiences of organisations in the EU using AI. Similarly, the research did not consider extra European contexts, even though the extra-territoriality of the GDPR also creates an obligation on some organisations not located in the EU. Moreover, it did not consider other countries which created DP and privacy legislation directly influenced by the GDPR.

The specific UK focus of the research means that its findings can have some limitations in terms of their generalisability and applicability to different geographical contexts where other socio-political, business, and legislative systems may foster different practices. Therefore, further research will be needed to understand those specificities. However, *CRAIDA* and the two models are not dependent on specific legislation and can therefore be used instrumentally — and with caution — in other contexts to help to identify their peculiarities and the specific needs related to the implementation and use of AI.

Therefore, even if the study is based on a comprehensive analysis of the UK context only, the researcher is confident that the findings and the strategic suggestions can help stimulate and guide scientific studies in other national or transnational contexts.

b. AI maturity and GDPR implementation

This limitation originates from the level of awareness and use of AI and GDPR found within UK organisations. The researcher was expecting to explore the praxis of organisations in more mature stages. Specifically, the expectation was to investigate the implementation of Art. 22 GDPR by organisations making decisions via automated systems. That could have provided the opportunity to analyse how they were meeting the specific requirements, such as the right to human intervention and the right to information/explanation. However, the encountered level of maturity of the implementation of AI and GDPR in the case studies could fulfil this expectation only partially, while some crucial information around Art. 22 was provided by the experts.

c. Limitations from the methodology

There are also some limitations arising from the specific case study approach.

- The amount and degree of detail of data collected made the analysis at times challenging, and it was not easy to represent the high level of complexity in an accessible style. These difficulties were overcome using pragmatic rationality in selecting and structuring the data and in analysing and presenting some of the data for conference audiences.

- Due to the small sample of organisations (two organisations only), there is the risk that the findings could not be easily generalised. Gathering data from the group of experts and drawing on their rich knowledge of the sector and analysing its findings with those from the case studies, helped hugely to reduce this specific risk. Some of the challenges identified in the case studies had also been previously encountered in a broad range of organisations as was evidenced by the initial study with the experts. Therefore, it is possible to expect that the findings would also be plausible in other similar GDPR regulated environments.

- The knowledge, and professional and personal experience of the researcher were important in shaping, conducting, and making decisions during the research (e.g., prioritising some topics, for example, power). Even though such decisions were made around data and its issues, a constant rigorous judgement was applied throughout the journey in providing and presenting adequate evidence.

- The focus of the case studies was on the organisation. The end-user was out-of-scope for this research.

Despite these minor limitations, this research presented a robust and reflective methodology, leading to novel insights in the field.

d. Expectations around decision-making processes

The researcher was expecting to find more data on decision-making processes (e.g., trade-offs) and potential conflicts faced by individuals and boards in establishing strategies and in making decisions around specific cases. The intent was to explore how ethical dilemmas emerged, were interpreted, and solved at different levels and in different settings. This was only partially fulfilled. While the amount of data provided information on some issues and doubts faced by participants, this did not provide the deep insights around the more philosophical dilemmas in decision-making processes as hoped for by the researcher.

9.6 Future research

As the topics of the research are evolving rapidly, further research is needed on various aspects. The three most urgent areas are related to Augmented AI and the relationship with external stakeholders.

a. Management of Augmented AI models

The role, power, and understanding of the human in the loop, with some psychological aspects in the interaction with ML, should be carefully researched, for example, around bias awareness and acquisition of knowledge. The assumed level of risk placed on this model is usually low or non-existent. This is per se a high risk for organisations underestimating AI and overestimating human intervention.

The need to better understand the complexity of this decision-making process is also necessary for those organisations willing to raise awareness on active participation, responsibility, and accountability in the “making” (the full innovation cycle).

Furthermore, specific training for those interacting with AI/ML and using predictions to make decisions should be created and regulated.

b. Regulating decisions made in Augmented AI models

While Art. 22 regulates decisions made via automated systems, the decisions made by humans in augmented systems are not regulated. Such decisions are not simply human decisions resulting only from human judgement. They are the result of a complex process where the knowledge embedded in data, algorithms and predictions encounters the knowledge in humans. The prediction becomes the foundation for human judgement and decisions. The responsibilities associated with both parts of the process, ML and Human, should be better researched and regulated.

c. Vendor management, multiparty liability, and Augmented AI models.

External stakeholders that build AI systems should bear part of the responsibility for decisions made using those systems. While issues around accountability in DP are evolving, similar issues around contractual responsibility and multiparty liability in multiparty ecosystems need to be researched and regulated considering AI/ML peculiarities.

Therefore, the issues and risks associated with the creation and use of AI systems supporting human decisions deserve more and urgent attention.

9.7 Reflections on the research journey

This research has been an incredible, exciting, difficult, and privileged journey. Some aspects of this journey have, more than others, shaped my understanding and approach.

Theoretical framework. The identification of the most appropriate theoretical framework created some difficulties. This was mainly due to the multidisciplinary character of the research which includes law, technology, management, and ethics.

Additionally, the dominant debate around AI and ethics felt quite unsatisfactory. It lacked the practical dimension, and its avoidance to look at challenges faced by those “doing” AI and GDPR ultimately felt deeply unfair. While the general focus was on “bad” companies using AI to exploit personal data, I was looking for a different approach that did not take for granted the willingness of people in organisations to exploit data. An approach that could consider other less visible aspects, such as individual agency, the desire to do good, and the performativity linked to the role. The desire to explore the experience of individuals and organisational praxis added further complexity to the research. I needed a holistic theory capable of uncovering the complexity that I suspected structured the field, and which I was hoping to unravel.

The turning point was a talk on RRI, IS and ethics given by Bernd Stahl at the Academy for Information Systems (UKAIS) in 2018. That talk gave me a sense of direction and it provided elements that would later form my theoretical framework. The choice of Feenberg and CTT was inspired by political, sociological, and philosophical studies conducted prior to the PhD. Discussions around power, oppression, resistance, and the role of culture were not new. The individual as an active agent of change in CTT was exactly what I needed for my theoretical framework.

Other potentially interesting aspects being briefly considered (again from former studies), were some elements that originated within the Gestalt tradition, such as the awareness of biases via practice. Although believing in the potential of the idea, that was discharged due to time constraints.

Intuitions. The exploration of AI and DP was the evolution of research carried out for my M.Sc. dissertation. The M.Sc. in Information Systems Management done prior to the Ph.D. provided a strong foundation for this research, plus a published research paper, and an invaluable supervisor who later became my PhD supervisor. Choosing my supervisor was my first lucky intuition. The topic of the dissertation, emerging technology and the GDPR, was chosen in 2016 when the Regulation was mainly unknown. That choice turned out to be another good intuition.

The M.Sc. dissertation set out the direction for future research and created the desire to further explore DP and AI. The possibility came with the *Pathway to Excellence Research Studentship* awarded by the University of Salford which made my desire possible.

Choosing AI and DP was another fortunate intuition. When I started my research in 2017 not many were aware of the peculiarities, potential, and risks of AI. ML had a revival the year before and organisations were becoming interested in AI to reduce cost and boost productivity. Around the same time, Brexit happened, with all the following chaos and huge amount of uncertainty, and the scandal involving Cambridge Analytica which gave visibility to the risks of AI. Subsequently, awareness of the importance of DP exploded.

More and more people and organisations are now working on AI ethics. And yet, most of them are still not looking at what is really happening inside organisations, preferring to discuss principles and obligations which risk becoming inapplicable and irrelevant. AI ethics is trendy, and the risk of "ethics washing" done with superficial promises is also real. Additionally, the overestimation of Augmented AI as a solution to AI risks is deeply disturbing, and its peculiarities need to be thoroughly researched as soon as possible.

Personal experiences. Some personal experiences have impacted my research in different ways. Encountering a "human in the loop" while travelling at an airport was an important moment in my understanding of decision-making processes and the role of humans.

While coming back to the UK, the facial recognition of the automated security control system did not recognise my face. I was then invited to proceed to the security checks performed by humans. The first officer appeared uncertain and kept looking at my face and my passport. Unable to decide if the person in the picture was really me, he asked a second officer for some help to verify my identity. Again, the second one started checking my passport and then staring at my face. I suspect neither my uncomfortable feelings experienced in those moments, nor the sense of complete powerlessness amplified by Brexit helped my body language to transmit reassuring signals. The second human in the loop could not decide either. I was then invited to have another picture taken by another machine. After another check, I was finally let go. Four different stages to validate my identity. My face had changed due to some health issues, not a lot but enough to have my identity questioned. Both AI and two humans in the loop could not easily decide.

I am very aware of my privileged status. I am an EU citizen, and I am a white person (although this may be questionable for some). These are not small details in the context of border control and Brexit. Other less privileged people are not always given those extra checks and time. This shows that fair and regulated processes need to be established also for decision making involving humans.

Finally, some health issues experienced since the first year of my PhD have deeply impacted my research. A pandemic on top of them has not helped either. Both have slowed down my research. Both have increased the sense of my otherness, the need for more protection in the face of AI and my resilience.

9.8 Conclusion

This chapter presented the conclusion of this research thesis. After evaluating the research according to its aims and objectives, the chapter presented the contributions of the research to theory, knowledge, and practice, and its limitations (e.g., methodological choices). It then provided suggestions for future research, highlighting the need to explore the management of augmented AI and multi-stakeholder AI governance.

The chapter finally ends with some reflections, describing how some academic and personal events shaped the whole research journey.

REFERENCES

ACM FAccT Conference. (2021). *FAccT Conference*. Retrieved 22 August 2021, from https://facctconference.org

Adams, J., Khan, H. T. A., Raeside, R., & White, D. I. (2007). *Research methods for graduate business and social science students*. SAGE publications India.

Addis, C., & Kutar, M. S. (2018, March). The general data protection regulation (GDPR), emerging technologies and UK organisations: awareness, implementation and readiness. In *UK Academy for Information Systems Conference Proceedings 2018* (p. 29). UKAIS–UK Academy for Information Systems.

Adobe Communications Team. (2016). *Let's Get Experimental: Behind the Adobe MAX Sneaks*. Retrieved 21 August 2021, from https://blog.adobe.com/en/publish/2016/11/04/lets-get-experimental-behind-the-adobe-max-sneaks.html

Affectiva. (2018). *Emotion AI Overview What is it and how does it work?* Affectiva Website. Retrieved 1 July 2021, from https://www.affectiva.com/emotion-ai-overview/

Affectiva. (2021). *Our Evolution from Emotion AI to Human Perception AI.* 2021. Retrieved 21 August 2021, from https://blog.affectiva.com/our-evolution-from-emotion-ai-to-human-perception-ai

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: the simple economics of artificial intelligence*. Harvard Business Press.

Agrawal, A., Gans, J., & Goldfarb, A. (2019). *An economic perspective on Artificial Intelligence*. *NATO Defence College*, 7-15.

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human Trajectory Prediction in Crowded Spaces. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 961–971. https://doi.org/10.1109/CVPR.2016.110

Allcott, H., & Gentzkow, M. (2017). Social Media and Fake News in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211–236. https://doi.org/10.1257/jep.31.2.211

Allen, G., & Chan, T. (2017). *Artificial Intelligence and National Security*. Retrieved 22 August 2021, from https://www.belfercenter.org/sites/default/files/files/publication/AI NatSec - final.pdf

Alvesson, M., & Deetz, S. (2000). *Doing critical management research*. Sage.

Amazon. (2018). *Amazon Prime Air*. Retrieved 21 August 2021, from https://www.amazon.com/Amazon-Prime-Air/b?ie=UTF8&node=8037720011

Ammanath, B., Hupfer, S., & Jarvis, D. (2020). Thriving in the era of pervasive AI: Deloitte's state of AI in the enterprise. *Deloitte Insights*.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. Retrieved 21 August 2021, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Angwin, J., Varner, M., & Ariana, T. (2017). *Facebook Enabled Advertisers to Reach 'Jew Haters.'*. Retrieved 21 August 2021, from *https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters*

Archer, M., Bhaskar, R., Collier, A., Lawson, T., & Norrie, A. (2013). *Critical realism: Essential readings*. Routledge.

Areheart, B. A., & Roberts, J. L. (2019). GINA, Big Data, and the Future of Employee Privacy.(Genetic Information Nondiscrimination Act of 2008). *Yale Law Journal*, *128*(3), 710.

Arik, S. O., Chen, J., Peng, K., Ping, W., & Zhou, Y. (2018). *Neural Voice Cloning with a Few Samples*. arXiv preprint arXiv:1802.06006.

Arner, D. W., Barberis, J. N., & Buckley, R. P. (2016). *The emergence of regtech 2.0: from know your customer to know your data*. Journal of Financial Transformation, vol. 44, pages 79-86.

Arthur, W. B. (2017). Where is technology taking the economy? *McKinsey Quarterly*, 697.

Article 29 Data Protection Working Party. (2018). *Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*. Retrieved 21 August 2021, from http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053

Article 29 Data Protection Working Party WP 223. (2018). *Article 29 Working Party Guidelines on consent under Regulation 2016/679. Adopted on 28 November 2017 As last Revised and Adopted on 10 April 2018*. The European Union. Retrieved 21 August 2021, from http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=623051

Asencio-Cortes, G., Florido, E., Troncoso, A., & Martínez-Álvarez, F. (2016). A novel methodology to predict urban traffic congestion with ensemble learning. *Soft Computing*, *20*(11), 4205–4216. https://doi.org/10.1007/s00500-016-2288-6

Barocas, S., & Selbst, A. (2016a). Big Data 's Disparate Impact. *California Law Review*, *104*(1), 671–729. https://doi.org/http://dx.doi.org/10.15779/Z38BG31

Bass, J. M., Beecham, S., & Noll, J. (2018). Experience of industry case studies : a comparison of multi case and embedded case study methods. In *Proceedings of the 6th International Workshop on Conducting Empirical Studies in Industry* (pp. 13-20).13–20. https://doi.org/10.1145/3193965.3193967

Baum, S. (2017). A survey of artificial general intelligence projects for ethics, risk, and policy. *Global Catastrophic Risk Institute Working Paper*, 11–17.

BBC Technology. (2016). *Adobe Voco "Photoshop-for-voice" causes concern*. BBC News. Retrieved 21 August 2021, from http://www.bbc.co.uk/news/technology-37899902

Beavers, A. F. (2002). Phenomenology and artificial intelligence. *Metaphilosophy*, *33*(1-2), 70–82.

Bell, E., Bryman, A., & Harley, B. (2018). *Business research methods*. Oxford university press.

Bellman, R. (1978). *An introduction to artificial intelligence: can computer think?* (No. 04; Q335, B4.).

Benbasat, I., Goldstein, D. K., & Mead, M. (1987). The Case Research Strategy in Studies of Information Systems. *MIS Quarterly*, *11*(3), 369. https://doi.org/10.2307/248684

Benjamin, R. (2019). Assessing risk, automating racism. *Science*, *366*(6464), 421–422.

Bergstein, B. (2019). From Intelligent Systems to Intelligent Organizations. *Research-Technology Management*, *62*(3), 31–37. https://doi.org/10.1080/08956308.2019.1587300

Berkeley, M. I. R. I.-. (2019). *Machine Intelligence Research Institute*. Retrieved 21 August 2021, from https://intelligence.org/

Beynon-Davies, P. (2011). The UK national identity card. *Journal of Information Technology Teaching Cases*, *1*(1), 12–21.

Bhaskar, R. (2013). *A realist theory of science*. Routledge.

Binns, R. (2017). *Fairness in Machine Learning: Lessons from Political Philosophy*. *2016*, 1–11. http://arxiv.org/abs/1712.03586

Bird, S., Barocas, S., Diaz, F., Crawford, K., & Wallach, H. (2016). 08_Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI. *Workshop on Fairness, Accountability, and Transparency in Machine Learning*. https://ssrn.com/abstract=2846909

Blumer, H. (1986). *Symbolic interactionism: Perspective and method*. Univ of California Press.

Borgesius, F. J. Z. (2018). *Discrimination, artificial intelligence, and algorithmic decision-*

making. Retrieved 21 August 2021, from https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73

Bosworth, F., Graham, P., Hache, A., Pellizzer, V., Jansen, F., Pagano, F., Toupin, S., Vergés, N., York, J., Van Dessel, M., & Waters, C. (2015). *Zen and the art of making tech work for you*. Tactical Technology Collective. Retrieved 21 August 2021, from https://gendersec.tacticaltech.org/wiki/index.php/Complete_manual

Bowling, A. (2002). Research Methods in Health Care. *Open University Press, Buckingham*.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brey, P., Lundgren, B., Macnish, K., Ryan, M., Lundgren, B., Macnish, K., & Ryan, M. (2019). *Guidelines for the Ethical Use of AI and Big Data Systems. Shaping the ethical dimensions of smart information systems– a European perspective (SHERPA)*. Retrieved 21 August 2021, from https://www.project-sherpa.eu/wp-content/uploads/2019/12/use-final.pdf

Brodsky, L., & Oakes, L. (2017). Data sharing and open banking. *McKinsey on Payments July*. Retrieved 23 August 2021, from https://www.mckinsey.com/industries/financial-services/our-insights/data-sharing-and-open-banking

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *ArXiv Preprint ArXiv:2005.14165*.

Brundage, M., Avin, S., Clark, J., Allen, G. C., Flynn, C., Farquhar, S., Crootof, R., & Bryson, J. (2018). *The Malicious Use of Artificial Intelligence : Forecasting , Prevention , and Mitigation*. *February*, 99. Retrieved 23 August 2021, from https://maliciousaireport.com

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530–1534.

Bughin, J., Catlin, T., Hirt, M., & Willmott, P. (2018). Why digital strategies fail. *McKinsey Quarterly*. Retrieved 21 August 2021, from https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/why-digital-strategies-fail

Burr, V. (2015). *Social constructionism*. Routledge.

Butler, J. (2011). *Bodies that matter: On the discursive limits of sex*. Routledge.

Butterworth, M. (2018). The ICO and artificial intelligence: The role of fairness in the GDPR framework. *Computer Law and Security Review*, *34*(2), 257–268. https://doi.org/10.1016/j.clsr.2018.01.004

California Consumer Privacy Act of 2018 CCPA, (2018) (testimony of California State Legislature). Retrieved 22 August 2021, from https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375

Caliskan-Islam, A., Yamaguchi, F., Dauber, E., Harang, R., Rieck, K., Greenstadt, R., & Narayanan, A. (2015). When coding style survives compilation: De-anonymizing programmers from executable binaries. *ArXiv Preprint ArXiv:1512.08546*.

Callahan, A., & Shah, N. H. (2017). *Machine Learning in Healthcare*. https://doi.org/10.1016/B978-0-12-809523-2.00019-4

Calo, R. (2017). Artificial Intelligence policy: a primer and roadmap. *UCDL Rev.*, *51*, 399.

Campbell, M., Hoane, A. J., & Hsu, F. (2002). Deep Blue. *Artificial Intelligence*, *134*(1–2), 57–83.

Campolo, A., Sanfilippo, M., Whittaker, M., & Kate Crawford. (2017). *AI Now 2017 Report*. Retrieved 21 August 2021, from https://ainowinstitute.org/AI_Now_2017_Report.pdf

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare : Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, 1721–1730. https://doi.org/10.1145/2783258.2788613

Case, C. (2014). 131/12, Google Spain SL, Google Inc. v Agencia Española de Protección de Datos (AEPD), Mario Costeja González. Retrieved 21 August 2021, from https://curia.europa.eu/juris/document/document.jsf?docid=152065&doclang=EN

Casey, B., Farhangi, A., & Vogl, R. (2018). *Rethinking Explainable Machines: The GDPR's' Right to Explanation'Debate and the Rise of Algorithmic Audits in Enterprise*. *Berkeley Tech. LJ*, *34*, 143.

Cesaratto, B. G. (2019). *Washington State Considers Comprehensive Data Privacy Act to Protect Personal Information*. The Nationa Law Review. Retrieved 21 August 2021, from https://www.natlawreview.com/article/washington-state-considers-comprehensive-data-privacy-act-to-protect-personal

Chen, A. (2019). The guy who made a tool to track women in porn videos is sorry. *MIT Technology Review*. Retrieved 21 August 2021, from https://www.technologyreview.com/s/613607/facial-recognition-porn-database-privacy-gdpr-data-collection-policy/

Chen, W., & Hirschheim, R. (2004). A paradigmatic and methodological examination of information systems research from 1991 to 2001. *Information systems journal*, 14(3), 197-235.

Chesney, R., & Citron, D. K. (2018). *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. Calif. L. Rev.*, *107*, 1753.

Chiou, S., Music, C., Sprague, K., & Wahba, R. (2001). A Marriage of Convenience: The Founding of the MIT Artificial Intelligence Laboratory. *Structure of Engineering Revolutions*.

Chokshi, N. (2018). Is Alexa Listening? Amazon Echo Sent Out Recording of Couple's Conversation. *The New York Times*. Retrieved 21 August 2021, from https://www.nytimes.com/2018/05/25/business/amazon-alexa-conversation-shared-echo.html

Christl, W. (2017). *Corporate Surveillance in Everyday Life: How Companies Collect, Combine, Analyze, Trade, and Use Personal Data on Billions.* Cracked Labs, Vienna, Austria. Retrieved 23 August 2021, from https://blog.fdik.org/2017-10/CrackedLabs_Christl_CorporateSurveillance.pdf

Chung, H., Iorga, M., Voas, J., & Lee, S. (2017). Alexa, can I trust you? *Computer*, *50*(9), 100–104.

CIFAR Canadian Institute for Advanced Research. (2017). *Pan-Canadian Artificial Intelligence Strategy Overview*. Retrieved 21 August 2021, from https://www.cifar.ca/assets/pan-canadian-artificial-intelligence-strategy-overview/

Cimpanu, C. (2019). *Hackers breach 62 US colleges by exploiting ERP vulnerability*. ZDNet. Retrieved 21 August 2021, from https://www.zdnet.com/article/hackers-breach-62-us-colleges-by-exploiting-erp-vulnerability

Clifford, D., & Ausloos, J. (2017). *Data Protection and the Role of Fairness. Yearbook of European Law*, *37*, 130-187

CognitionX. (2018a). *CognitionX - The AI advice platform - Case Studies*. CognitionX. Retrieved 13 November 2018, from https://cognitionx.com/directory/casestudies/

CognitionX. (2018b). *CognitionX - The AI advice platform - Schneider Electric + Recruitment*. 2018. Retrieved 13 November 2018, https://cognitionx.com/directory/casestudies/7/Schneider Electric %2B Recruitment

Cole, S. (2017). *Facial Recognition for Porn Stars Is a Privacy Nightmare Waiting to Happen*. Motherboard. Retrieved 21 August 2021, from https://motherboard.vice.com/en_us/article/a3kmpb/facial-recognition-for-porn-stars-is-a-privacy-nightmare-waiting-to-happen

Collier, A. (1999). *Being and worth*. London.

Collins, B. (2021). Microsoft Could Bring You Back From The Dead... As A Chat Bot. *Forbes*. Retrieved 21 August 2021, from https://www.forbes.com/sites/barrycollins/2021/01/04/microsoft-could-bring-you-back-from-the-dead-as-a-chat-bot/?sh=4515bfc85f70

Concordia. (2016). *Cambridge Analytica - The Power of Big Data and Psychographics*. Retrieved 21 August 2021, from https://www.youtube.com/watch?time_continue=356&v=n8Dd5aVXLCc

Corfield, G. (2019). *Lancaster Uni data breach hits at least 12,500 wannabe students*. The Register. Retrieved 21 August 2021, from https://www.theregister.co.uk/2019/07/23/lancaster_university_data_breach/

The European convention on human rights (ECHR), (1952) (testimony of Council of Europe). Retrieved 21 August 2021, from https://www.echr.coe.int/Documents/Convention_ENG.pdf

Council of Europe. (1981). Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data (108). *European Treaty Series No108. Council of Europe: Strasbourg*. Retrieved 23 August 2021.https://rm.coe.int/1680078b37

Council of Europe. (2017). *Algorithms and Human Rights. Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications*. Retrieved 23 August 2021, from https://edoc.coe.int/en/internet/7589-algorithms-and-human-rights-study-on-the-human-rights-dimensions-of-automated-data-processing-techniques-and-possible-regulatory-implications.html

Amendments to Convention 108, (2018) (testimony of Council of Europe). Retrieved 21 August 2021, from https://www.coe.int/en/web/data-protection/convention108/amendments

Cover, R. (2015). *Digital identities: Creating and communicating the online self.* Academic Press.

Crawford, K. (2016). Artificial intelligence’s white guy problem. *The New York Times*. Retrieved 21 August 2021, from https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

Crawford, K. (2017). *The Trouble with Bias - NIPS 2017 Keynote - Kate Crawford #NIPS2017*. The Artificial Intelligence Channel. Retrieved 21 August 2021, from https://www.youtube.com/watch?v=fMym_BKWQzk

Crawford, K. (2018). *AI at Google. Our Principles*. Twitter. https://twitter.com/katecrawford/status/1005127257248161792

Crawford, K., & Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *BCL Rev.*, *55*, 93.

Creswell, J. W. (2014). *A concise introduction to mixed methods research*. Sage Publications.

Crotty, M. (1998). *The foundations of social research : meaning and perspective in the research process*. London : SAGE.

Davies, B., & Wendes, S. (2014). *Technology and Innovation for the Future of Composites Manufacturing*. Retrieved 23 August 2021, from https://studylib.net/doc/10747759/technology-and-innovation-for-the-future-of-composites-ma...

Day, M., Turner, G., & Drozdiak, N. (2019). Amazon Workers Are Listening to What You Tell Alexa. *Bloomberg*. Retrieved 21 August 2021, from https://www.bloomberg.com/news/articles/2019-04-10/is-anyone-listening-to-you-on-alexa-a-global-team-reviews-audio

de Certeau, M. (1984). The practice of everyday life, trans. Steven Rendall.Berkeley: University of California Press.

DeepMind. (2021). *DeepMind Who we are*. Retrieved 21 August 2021, from https://deepmind.com/about#our_story

Delvaux, M. (2016). Draft report with recommendations to the Commission on Civil Law Rules on Robotics. *European Parliament: Brussels, Belgium*, 22.

Delvaux, Mady. (2017). REPORT with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103 (INL). *European Parliament*. Retrieved 21 August 2021, from http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+REPORT+A8-2017-0005+0+DOC+PDF+V0//EN

Deng, L., & Yu, D. (2014). Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, *7*(3–4), 197–387.

Denham, E. (2017). *Four lessons NHS Trusts can learn from the Royal Free case*. Ico.Org.Uk. Retrieved 21 August 2021, from https://ico.org.uk/about-the-ico/news-and-events/blog-four-lessons-nhs-trusts-can-learn-from-the-royal-free-case/

Di Ieva, A. (2019). AI-augmented multidisciplinary teams: hype or hope? *The Lancet*, *394*(10211), 1801.

Diebold, F. X. (2012). *On the Origin (s) and Development of the Term 'Big Data'* (No. 12-

037). Penn Institute for Economic Research, Department of Economics, University of Pennsylvania.

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer Nature.

Dilworth, J. (2017). *Advanced In-Database Analytics on the GPU*. Kinetica. Retrieved 21 August 2021, from https://www.kinetica.com/blog/in-database-analytics-gpu/

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S. J., O'Brien, D., Shieber, S., Waldo, J., Weinberger, D., & Wood, A. (2017). Accountability of AI Under the Law: The Role of Explanation. *Ssrn*, *December 2016*. *arXiv preprint arXiv:1711.01134*.

Dr Murphy. (2019). *The #BadBotFiles This open-ended thread provides links to news stories & media coverage relating to the @babylonhealth Chatbot. These articles won't be found on the Babylon Website, but are recommended reading for anyone wishing to learn more about Babylo*. Twitter. Retrieved 21 August 2021, from https://twitter.com/DrMurphy11/status/1094286216705568768

Dresner, S. (2014). EU Data Protection Reform: Close to the finishing line. *European Data Protection Supervisor*. Retrieved 21 August 2021, from https://edps.europa.eu/sites/edp/files/publication/14-11-06_edps_bfdi_report_en.pdf

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580.

Dyche, J. (2018). *5 questions CEOs are asking about AI | CIO*. CIO from IDG Communication. Retrieved 21 August 2021, from https://www.cio.com/article/3318639/artificial-intelligence/5-questions-ceos-are-asking-about-ai.html

Ebers, M. (2019). *Regulating AI and robotics: Ethical and legal challenges*.

Edwards, L., & Veale, M. (2017). Slave to the Algorithm? Why a Right to Explanationn is Probably Not the Remedy You are Looking for. In *SSRN Electronic Journal* (Vol. 2017). https://doi.org/10.2139/ssrn.2972855

EFTA. (2018). *Incorporation of the GDPR into the EEA Agreement*. Retrieved 21 August 2021, from https://www.efta.int/EEA/news/Incorporation-GDPR-EEA-Agreement-508041

Ekowo, M., & Palmer, I. (2016). The Promise and Peril of Predictive Analytics in Higher Education: A Landscape Analysis. *New America*.

Elliot, M., O'Hara, K., Raab, C., O'Keefe, C. M., Mackey, E., Dibben, C., Gowans, H., Purdam, K., & McCullagh, K. (2018). Functional anonymisation: Personal data and the data environment. *Computer Law and Security Review*, *34*(2), 204–221.

Erdos, D. (2015). The Emergence of Personal Data Protection as a Fundamental Right of the EU. By Gloria González Fuster [Cham: Springer, 2014. xvi, 274 pp. ISBN 3319050222.]. *The Cambridge Law Journal*, *74*(2), 374–375. doi:10.1017/S0008197315000276

EU Directorate-General for Research and Innovation. (2018). *Smart lie-detection system to tighten EU's busy borders*. Retrieved 21 August 2021, from https://ec.europa.eu/research-and-innovation/en/projects/success-stories/all/smart-lie-detection-system-tighten-eus-busy-borders

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

European Commission. (2020a). *Responsible research & innovation*. Horizon 2020. Retrieved 21 August 2021, from https://ec.europa.eu/programmes/horizon2020/en/h2020-section/responsible-research-innovation

European Commission. (2020b). *eIDAS Regulation*. European Commission - Policies. Retrieved 26 August 2021, from https://digital-strategy.ec.europa.eu/en/policies/eidas-regulation

European Commission. (2021). *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, (2021) (testimony of European Commission). Retrieved 21 August 2021, from https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence

European Council. (1981). *Convention for the protection of individuals with regard to automatic processing of personal data* (Vol. 108). Council of Europe. Retrieved 21 August 2021, from https://rm.coe.int/1680078b37

European Data Protection Board. (2019). *Statement 2/2019 on the use of personal data in the course of political campaigns*. Retrieved 21 August 2021, from https://edpb.europa.eu/sites/edpb/files/files/file1/edpb-2019-03-13-statement-on-elections_en.pdf

European Data Protection Supervisor. (2016). Artificial Intelligence, Robotis, Privacy and Data Protection. *Room Document for the 38th International Conference of Data Protection and Privacy Commissioners*. Retrieved 21 August 2021, from https://edps.europa.eu/data-protection/our-work/publications/other-documents/artificial-intelligence-robotics-privacy-and_en

European Institute for Theoretical Neuroscience (EITN). (2018). *Understanding Cognition*. The Human Brain Project. Retrieved 21 August 2021, from https://www.humanbrainproject.eu/en/

Directive (EU) 2015/2366 f the European Parliament and of the Coulcil of 25 November 2015 on payment services in the internal market, amending Directives 2002/65/EC, 2009/110/EC and 2013/36/EU and Regulation (EU) No 1093/2010, and repealing Directive 200, (2015) (testimony of European Parliament and of the Council). https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32015L2366

Directive (EU) 2015/849 of the European Parliament and of the Council of 20 May 2015 on the prevention of the use of the financial system for the purposes of money laundering or terrorist financing, amending Regulation (EU) No 648/2012 of the European Par, (2015) (testimony of European Parliament and of the Council). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32015L0849

Directive (EU) 2018/843 of the European Parliament and of the Council of 30 May 2018 amending Directive (EU) 2015/849 on the prevention of the use of the financial system for the purposes of money laundering or terrorist financing, and amending Directives, (2018) (testimony of European Parliament and of the Council). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32018L0843

European Union. (2007). *Treaty of Lisbon: Amending the Treaty on European Union and the Treaty establishing the European Community*. Office for Official Publications of the European Community. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A12007L%2FTXT

European Union. (2019). *EU treaties*. https://europa.eu/european-union/law/treaties_en

European Union Agency for Fundamental Rights. (2018). *European Union Agency for Fundamental Rights*. Retrieved 26 August 2021, https://fra.europa.eu/en/about-fra

European Union Agency for Fundamental Rights. (2018). *Handbook on European data protection law - 2018 edition*. Retrieved 23 August 2021, from https://fra.europa.eu/sites/default/files/fra_uploads/fra-coe-edps-2018-handbook-data-protection_en.pdf

Eyal, N. (2014). *Hooked: How to build habit-forming products*. Penguin.

FAT/ML Conference. (2018). *FAT/ML Conference - 2018 Papers*. Retrieved 23 August 2021, from https://www.fatml.org/schedule/2018/page/papers-2018

FAT Conference. (2018). Retrieved 23 August 2021, from *FAT Conference*.

https://fatconference.org/

Feenberg, A. (2002). *Transforming technology: A critical theory revisited*. Oxford University Press.

Feenberg, A. (2005). Critical theory of technology: An Overview. *Tailoring Biotechnologies*, *1*(1), 47–64. Retrieved 23 August 2021, from https://www.sfu.ca/~andrewf/books/critbio.pdf

Feenberg, A. (2010). *Between reason and experience: Essays in technology and modernity*. Mit Press.

Feenberg, A. (2015). Making the gestalt switch. *Postphenomenological Investigations*, 229. *Essays on Human-Technology Relations*, 229-36.

Feenberg, A. (2017). *Technosystem*. Harvard University Press.

Feher, K. (2019). Digital identity and the online self: Footprint strategies–An exploratory and comparative research study. *Journal of Information Science*, *47*(2), 192-205.

Fenton, A., Fletcher, G., & Griffiths, M. (2019). *Strategic Digital Transformation: a results-driven approach*. Routledge.

Fichman, R. G., Dos Santos, B. L., & Zheng, Z. E. (2014). Digital innovation as a fundamental and powerful concept in the information systems curriculum. *MIS Quarterly*, *38*(2). 329-A15.

Finlay, J., & Dix, A. (1996). *An introduction to artificial intelligence* (A. J. Dix (ed.)). London : UCL Press.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. *Berkman Klein Center Research Publication*, *2020–1*. Retrieved 24 August 2021, fromhttps://dash.harvard.edu/bitstream/handle/1/42160420/HLS%20White%20Paper%20Final_v3.pdf

Floridi, L. (2020). AI and Its new winter: From myths to realities. *Philosophy & Technology*, *33*(1), 1–3.

Fountaine, T., McCarthy, B., & Saleh, T. (2019). Building the AI-powered organization. *Harvard Business Review*, *97*(4), 62–73.

Fraser, H., Coiera, E., & Wong, D. (2018). Safety of patient-facing digital symptom checkers. *The Lancet*, *392*(10161), 2263–2264.

Friedler, S. A., & Wilson, C. (2018). Preface. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, *81*, 1–2. Retrieved 23 August 2021, from https://proceedings.mlr.press/v81/friedler18a.html.

Fu, Y. @yiqinfu. (2019). *A Germany-based Chinese programmer said he and some friends have identified 100k porn actresses from around the world, cross-referencing faces in porn videos with social media profile pictures. The goal is to help others check whether their girlfriends ev*. Twitter. Retrieved 26 August 2021, from https://archive.ph/VIXWc (original https://twitter.com/yiqinfu/status/1133215940936650754)

Fuster, G. G. (2020). *Artificial Intelligence and Law Enforcement - Impact on Fundamental Rights*. Retrieved 23 August 2021, from https://www.europarl.europa.eu/thinktank/en/document.html?reference=IPOL_STU(2020)656295

Future of Life Institute. (2017). *Superintelligence: Science or Fiction? | Elon Musk & Other Great Minds*. You Tube. Retrieved 22 August 2021, from https://www.youtube.com/watch?v=h0962biiZa4&t=3121s

Gadamer, H. G. (2006). Classical and philosophical hermeneutics. *Theory, Culture and Society*, *23*(1), 29–56. https://doi.org/10.1177/0263276406063228

Gao, J. (2014). Machine learning applications for data center optimization. *Google White Paper*, 1–13.

GCHQ. (2021). *Pioneering a New National Security. The Ethics of Artificial Intelligence*. Retrieved 23 August 2021, from https://www.gchq.gov.uk/files/GCHQAIPaper.pdf

Gebru, T., Hoffman, J., & Fei-Fei, L. (2017). *Fine-grained Recognition in the Wild: A Multi-Task Domain Adaptation Approach*. *1*. https://doi.org/10.1109/ICCV.2017.151

Gebru, T., Krause, J., Wang, Y., Chen, D., Deng, J., Aiden, E. L., & Fei-Fei, L. (2017). *Using Deep Learning and Google Street View to Estimate the Demographic Makeup of the US*. 1–41. https://doi.org/10.1073/pnas.1700035114

Gershgorn, D. (2017). *AI researchers are trying to combat how AI can be used to lie and deceive*. Retrieved 22 August 2021, from https://qz.com/1150240/ai-researchers-are-trying-to-combat-how-ai-can-be-used-to-lie-and-deceive/

Gershman, J. (2017, April 25). Imitation Game: The Legal Implications of Voice Cloning. *The Wall Street Journal*. Retrieved 23 August 2021, from https://blogs.wsj.com/law/2017/04/25/imitation-game-the-legal-implications-of-voice-cloning/

Gibbs, G. (2012). *Types of Case Study. You Tube*. Retrieved 22 August 2021, from https://www.youtube.com/watch?v=gQfoq7c4UE4&index=2&list=PLirEzjzoHKvyh3BOlR0Yady8NajHGhVcs&t=0s

Gligor, D. M., Pillai, K. G., & Golgeci, I. (2021). Theorizing the dark side of business-to-business relationships in the era of AI, big data, and blockchain. *Journal of Business Research*, *133*, 79–88.

Goldman Sachs. (2016). Profiles in Innovation: Artificial Intelligence-AI. *Machine Learning and Data Fuel the Future of Productivity*. Retrieved 23 August 2021, from https://www.gspublishing.com/content/research/en/reports/2019/09/04/a0d36f41-b16a-4788-9ac5-68ddbc941fa9.pdf

Goodfellow, I., Hwang, T., Goodman, B., & Rodriguez, M. (2017). *Machine Deception*. Nips Conferences. Retrieved 22 August 2021, from https://nips.cc/Conferences/2017/Schedule?showEvent=8763

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Networks*. 1–9. https://arxiv.org/abs/1406.2661

Goodfellow, I., McDaniel, P., & Papernot, N. (2018). Making machine learning robust against adversarial inputs. *Communications of the ACM*, *61*(7), 56–66.

Goodman, B., & Flaxman, S. (2016a). *European Union regulations on algorithmic decision-making and a "right to explanation.". AI magazine*, *38*(3), 50-57.

Goodman, B., & Flaxman, S. (2016b). *European Union regulations on algorithmic decision-making and a "right to explanation."* 1–47. https://doi.org/10.1609/aimag.v38i3.2741

Google. (2018). *AI at Google: our principles*. Google Website. Retrieved 22 August 2021, from https://blog.google/technology/ai/ai-principles/

Gov.UK. (1998). Data Protection Act 1998. *Legislation.Gov.Uk*. Retrieved 22 August 2021, from https://www.legislation.gov.uk/ukpga/1998/29/contents

Government Digital Service. (2020a). *GOV.UK Verify overview*. Retrieved 22 August 2021, from https://www.gov.uk/government/publications/introducing-govuk-verify/introducing-govuk-verify

Government Digital Service. (2020b). *Guidance GOV.UK Verify*. Retrieved 22 August 2021, from https://www.gov.uk/government/publications/introducing-govuk-verify/introducing-govuk-verify#who-can-use-govuk-verify

Graziano, M. (2016). *Why You Should Believe in the Digital Afterlife*. The Atlantic. Retrieved 22 August 2021, from https://www.theatlantic.com/science/archive/2016/07/what-a-digital-afterlife-would-be-like/491105/

Great Britain. Department for Business. (2021). *New strategy to unleash the transformational power of Artificial Intelligence*. Retrieved 22 August 2021, from https://www.gov.uk/government/news/new-strategy-to-unleash-the-transformational-power-of-artificial-intelligence

Great Britain. Department for Business, Energy and Industrial Strategy. (2017). *Industrial strategy: building a Britain fit for the future.*

Green, B., & Chen, Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 90–99. https://doi.org/10.1145/3287560.3287563

Greenfield, P. (2018). The Cambridge Analytica files: the story so far. *The Guardian*. Retrieved 22 August 2021, from https://www.theguardian.com/news/2018/mar/26/the-cambridge-analytica-files-the-story-so-far

Griffiths, M., Heinze, A., Fenton, A., & Fletcher, G. (2018, September). Digital business evolution: lessons from a decade of KTP industry projects. In *UK Academy for Information Systems Conference Proceedings 2018*. AIS eLibrary.

Guha, A., Grewal, D., Kopalle, P. K., Haenlein, M., Schneider, M. J., Jung, H., Moustafa, R., Hegde, D. R., & Hawkins, G. (2021). How artificial intelligence will affect the future of retailing. *Journal of Retailing*, *97*(1), 28–41.

Hacking, I. (1981). *Scientific revolutions*. Oxford University Press.

Hajian, S., & Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, *25*(7), 1445–1459. https://doi.org/10.1109/TKDE.2012.72

Hammersley, M. (2009). Why critical realism fails to justify critical social research. *Methodological Innovations Online*, *4*(2), 1–11.

Harbinja, E. (2017). Post-mortem privacy 2.0: theory, law, and technology. *International Review of Law, Computers and Technology*, *31*(1), 26–42. https://doi.org/10.1080/13600869.2017.1275116

Harvey, I. (2000). Robotics: Philosophy of mind using a screwdriver. *Evolutionary Robotics: From Intelligent Robots to Artificial Life*, *3*, 207–230.

Harwich, E., & Laycock, K. (2018). *Thinking on its own: AI in the NHS. Reform Research Trust.*

Hassabis, D., & Silver, D. (2017). *AlphaGo Zero: Learning from scratch*. DeepMind Publication. Retrieved 22 August 2021, from https://deepmind.com/blog/alphago-zero-learning-scratch/

Hassan, N. R., Mingers, J., & Stahl, B. (2018). Philosophy and information systems: where are we and where should we go? *European Journal of Information Systems*, *27*(3), 263–

277. https://doi.org/10.1080/0960085X.2018.1470776

Haugeland, J. (1989). *Artificial intelligence: The very idea*. MIT press.

Hazel, J. W., & Slobogin, C. (2018). Who Knows What, and When: A Survey of the Privacy Policies Proffered by US Direct-to-Consumer Genetic Testing Companies. *Cornell JL & Pub. Pol'y*, *28*, 35.

Helmholz, R. H. (1990). Continental Law and Common Law: Historical Strangers or Companions? *Duke Law Journal*, *1990*(6), 1207–1228.

Herriott, R. E., & Firestone, W. A. (1983). Multisite qualitative policy research: Optimizing description and generalizability. *Educational Researcher*, *12*(2), 14–19.

High-Level Expert Group on AI - European Commission. (2019). Retrieved 22 August 2021, from *Ethics guidelines for trustworthy AI.* https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hill, K. (2017). *How Facebook Outs Sex Workers*. Retrieved 22 August 2021, from https://gizmodo.com/how-facebook-outs-sex-workers-1818861596

Hill, K. and, & Mattu, S. (2018). *The House That Spied on Me*. Gizmodo. Retrieved 22 August 2021, from https://gizmodo.com/the-house-that-spied-on-me-1822429852

Hodson, H. (2019). DeepMind and Google: the battle to control artificial intelligence. *The Economist*. Retrieved 22 August 2021, from https://www.1843magazine.com/features/deepmind-and-google-the-battle-to-control-artificial-intelligence

Hornung, G., & Schnabel, C. (2009). Data protection in Germany I: The population census decision and the right to informational self-determination. *Computer Law & Security Review*, *25*(1), 84–88.

House of Lords. (2018). *Written evidence volume: AI in the UK: ready, willing and able?* Retrieved 22 August 2021, from https://www.parliament.uk/documents/lords-committees/Artificial-Intelligence/AI-Written-Evidence-Volume.pdf

Humanyze. (2018). *Humanyze - Analytics For Better Performance*. Humanyze. Retrieved 22 August 2021, from https://www.humanyze.com/

Hume, D. (2009). *A treatise of human nature : being an attempt to introduce the experimental method of reasoning into moral subjects*. Waiheke Island : Floating Press.

Iansiti, M., & Lakhani, K. R. (2020a). *Competing in the age of AI: strategy and leadership when algorithms and networks run the world*. Harvard Business Press.

Iansiti, M., & Lakhani, K. R. (2020b). From Disruption to Collision: The New Competitive Dynamics. *MIT Sloan Management Review*, *61*(3), 34–39.

IEEE Board. (2019). *Ethical Aspects of Autonomous and Intelligent Systems*. *June*, 1–4. Retrieved 22 August 2021, from https://globalpolicy.ieee.org/wp-content/uploads/2019/06/IEEE19002.pdf

IESE Business School. (2018). *10 Ways Artificial Intelligence Is Transforming Management*. Conference The Future of Management in an Artificial Intelligence-Based World. Retrieved 22 August 2021, from https://www.iese.edu/en/about-iese/news-media/news/2018/april/10-ways-artificial-intelligence-is-transforming-management

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth.* Indiana University Press

Information Commissioner's Office (ICO). (2015). *TalkTalk cyber attack – how the ICO's investigation unfolded*. Retrieved 22 August 2021, from https://ico.org.uk/about-the-ico/news-and-events/talktalk-cyber-attack-how-the-ico-investigation-unfolded/

Information Commissioner's Office (ICO). (2017). *Big Data, artificial intelligence, machine learning and data protection*. p6. Retrieved 22 August 2021, from https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

Information Commissioner's Office (ICO). (2018). *Consent*. ICO Website. Retrieved 22 August 2021, from https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/consent/

Information Commissioner's Office (ICO). (2020a). *Guidance on AI and data protection*. Retrieved 22 August 2021, from https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/guidance-on-ai-and-data-protection/#

Information Commissioner's Office (ICO). ICO. (2020b). *The Guide to the Sandbox.* Retrieved 22 August 2021, from https://ico.org.uk/for-organisations/regulatory-sandbox/the-guide-to-the-sandbox/

Information Commissioner's Office (ICO). ICO. (2020c). *What is the eIDAS Regulation?* ICO Website. Retrieved 22 August 2021, from https://ico.org.uk/for-organisations/guide-to-eidas/what-is-the-eidas-regulation/

Inglesant, P., Jirotka, M., & Hartswood, M. (2018). Responsible Innovation in Quantum Technologies applied to Defence and National Security. *NQIT*. Retrieved 22 August 2021, from https://nqit.ox.ac.uk/sites/www.nqit.ox.ac.uk/files/2018-11/Responsible Innovation in Quantum Technologies applied to Defence and National Security PDFNov18.pdf

ITU. (2017). *AI for GLOBAL* (Issue June). Retrieved 22 August 2021, from https://www.itu.int/en/ITU-T/AI/Documents/Report/AI_for_Good_Global_Summit_Report_2017.pdf

Jablon, R. (2021). University of California victim of nationwide hack attack. *Los Angeles Times*. Retrieved 22 August 2021, from https://www.latimes.com/world-nation/story/2021-04-02/university-of-california-victim-of-nationwide-hack-attack

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research*, *1*(2), 112–133.

Johnson, S. (2017). Invasive or Informative? Educators Discuss Pros and Cons of Learning Analytics. *Edsurge*. Retrieved 22 August 2021, from https://www.edsurge.com/news/2017-11-01-invasive-or-informative-educators-discuss-pros-and-cons-of-learning-analytics

Kaitlyn, T. (2019). Amazon is being sued for recording children’s voices with Alexa. *Vox*. Retrieved 22 August 2021, from https://www.vox.com/the-goods/2019/6/14/18679360/amazon-alexa-federal-lawsuit-child-voice-recording

Kamarinou, D., Millard, C., & Singh, J. (2016). *Machine Learning with Personal Data*. Retrieved 22 August 2021, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2865811

Kamiran, F., Žliobaite, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. In *Knowledge and Information Systems* (Vol. 35, Issue 3). https://doi.org/10.1007/s10115-012-0584-8

Kang, B., & Choo, H. (2016). A deep-learning-based emergency alert system. *ICT Express*, *2*(2), 67–70. https://doi.org/10.1016/j.icte.2016.05.001

Kang, H.-W., & Kang, H.-B. (2017). Prediction of crime occurrence from multi-modal data using deep learning. *Plos One*, *12*(4), e0176244.

Kay, M., Matuszek, C., & Munson, S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 3819–3828.

Kerry, C., Blythe, F., & Long, W. (2016). *How big will big data be under the GDPR?* The International Association of Privacy Professionals. Retrieved 22 August 2021, from https://iapp.org/

Ketokivi, M., & Mantere, S. (2010). Two strategies for inductive reasoning in organizational research. *Academy of management review*, *35*(2), 315-333. Retrieved 26 August 2021, from https://helda.helsinki.fi/dhanken/bitstream/handle/10138/45349/ketokivi-mantere_AMR.pdf?sequence=1

Kharif, O. (2014). Privacy fears over student data tracking lead to InBloom's shutdown. *Bloomberg Business*. Retrieved 23 August 2021, from https://www.bloomberg.com/news/articles/2014-05-01/inbloom-shuts-down-amid-privacy-fears-over-student-data-tracking

Khatchadourian, R. (2015). We Know How You Feel. *The New Yorker*. Retrieved 22 August 2021, from https://www.newyorker.com/magazine/2015/01/19/know-feel

Kind, C. (2021). *Containing the canary in the AI coalmine – the EU's efforts to regulate biometrics*. Ada Loverace Institute. Retrieved 22 August 2021, from https://www.adalovelaceinstitute.org/blog/canary-ai-coalmine-eu-regulate-biometrics/

Kloza, D., van Dijk, N., Gellert, R., Böröcz, I., Tanas, A., Mantovani, E., & Quinn, P. (2017). *Data protection impact assessments in the European Union: complementing the new legal framework towards a more robust protection of individuals*. Retrieved 23 August 2021, from https://virthost.vub.ac.be/LSTS/dpialab/images/dpialabcontent/dpialab_pb2017-1_final.pdf

Knight, W. (2016). *AI Winter Isn't Coming*. MIT Technology Review. Retrieved 22 August 2021, from https://www.technologyreview.com/s/603062/ai-winter-isnt-coming/

Koene, A., Clifton, C., Hatada, Y., Webb, H., & Richardson, R. (n.d.). *A governance framework for algorithmic accountability and transparency*. Retrieved 23 August 2021, from https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_STU(2019)624262

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, *110*(15), 5802–5805. https://doi.org/10.1073/pnas.1218772110

Kreps, D. (2019). *Understanding digital events: Bergson, Whitehead, and the experience of the digital*. Routledge.

Kreps, D., Blaynee, J., Kutar, M., & Griffiths, M. (2016). KTP and RRI: the perfect match. *ACM SIGCAS Computers and Society*, *45*(3), 332–336.

Kuchler, H. (2018). Facebook returns to facial recognition in Europe despite qualms. *Financial Times*. Retrieved 22 August 2021, from https://www.ft.com/content/7c978778-4297-11e8-803a-295c97e6fd0b

Kurshan, B. (2016). What EdTech Entrepreneurs Should Know And Do About Student Data Privacy. *Forces*. Retrieved 22 August 2021, from https://www.forbes.com/sites/barbarakurshan/2016/12/08/what-edtech-entrepreneurs-should-know-and-do-about-student-data-privacy/#58df06444dc0

Kurshan, B. (2017). The Elephant in the Room With EdTech Data Privacy. *Forbes*. Retrieved 22 August 2021, from https://www.forbes.com/sites/barbarakurshan/2017/06/22/the-elephant-in-the-room-with-edtech-data-privacy/#650f70d557a5

Kurzweil, R., Richter, R., Kurzweil, R., & Schneider, M. L. (1990). The age of intelligent machines (Vol. 579). MIT press Cambridge.

Kvasny, L., & Richardson, H. (2006). Critical research in information systems: looking forward, looking back. *Information Technology & People.*

Lamphere, C. (2018). *Does New Media Generation Technology Pose an Existential Threat to Factual Information ?* Retrieved 26 August 2021, from *https://www.infotoday.com/OnlineSearcher/Issue/7767-January-February-2018.shtml*

Langley, P. (2011). The changing science of machine learning. *Machine Learning*, *82*(3), 275–279. https://doi.org/10.1007/s10994-011-5242-y

Laverty, S. M. (2003). Hermeneutic Phenomenology and Phenomenology: A Comparison of Historical and Methodological Considerations. *International Journal of Qualitative Methods*, *2*(3), 21–35. https://doi.org/10.1177/160940690300200303

Lazar, J. (2017). *Research methods in human computer interaction* (J. H. Feng author & H. Hochheiser author (eds.); Second edi). Cambridge, Massachusetts : Morgan Kaufmann.

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, K., & Ha, N. (2018). AI platform to accelerate API economy and ecosystem. *International Conference on Information Networking*, *2018-Janua*(Ilsvrc 2012), 848–852. https://doi.org/10.1109/ICOIN.2018.8343242

Lee, L. (2018). *Alibaba to take on Kuala Lumpur's traffic in first foreign project*. Reuters. Retrieved 22 August 2021, from https://uk.reuters.com/article/us-alibaba-malaysia/alibaba-to-take-on-kuala-lumpurs-traffic-in-first-foreign-project-idUKKBN1FI0QV

Lee, P. (2018). *Every vendor wants to be… a data controller?! - Privacy, Security and Information Law Fieldfisher*. Fieldfisher Privacy Law Blog. Retrieved 22 August 2021, from https://privacylawblog.fieldfisher.com/2018/every-vendor-wants-to-be-a-data-controller

European Union (Withdrawal) Act 2018, (2018) (testimony of legislation.gov.uk.). Retrieved 22 August 2021, from https://www.legislation.gov.uk/ukpga/2018/16/contents/enacted/data.htm

Legislation UK. (2018). *Data Protection Act 2018*. Retrieved 22 August 2021, from https://www.legislation.gov.uk/ukpga/1998/29/contents

Lehman-Wilzig, S. N. (1981). Frankenstein unbound. Towards a legal definition of artificial intelligence. *Futures*, *13*(6), 442–457. https://doi.org/10.1016/0016-3287(81)90100-2

Lester, J. N., Cho, Y., & Lochmiller, C. R. (2020). Learning to do qualitative data analysis: A starting point. *Human Resource Development Review*, *19*(1), 94–106.

Lipton, P. (2003). *Inference to the best explanation*. Routledge.

Lohr, S. (2017). *How Do You Vote? 50 Million Google Images Give a Clue*. The New York Times. Retrieved 22 August 2021, from https://www.nytimes.com/2017/12/31/technology/google-images-voters.html

Lomas, N. (2021). *Orwellian' AI lie detector project challenged in EU court*. Retrieved 22 August 2021, from https://techcrunch.com/2021/02/05/orwellian-ai-lie-detector-project-challenged-in-eu-court/

Loucks, J., Davenport, T., & Schatsky, D. (2018). *State of AI in the Enterprise*. Retrieved 22 August 2021, from https://www2.deloitte.com/insights/us/en/focus/cognitive-technologies/state-of-ai-and-intelligent-automation-in-business-survey.html

Luca, M., Kleinberg, J., & Mullainathan, S. (2016). Algorithms need managers, too. *Harvard Business Review*, *94*(1), 20.

Lund, S., Madgavkar, A., Manyika, J., Smit, S., Ellingrud, K., Meaney, M., & Robinson, O. (2021). *The future of work after COVID-19*. McKinsey Global Institute, *18*.

Lynskey, O. (2014). Deconstructing Data Protection: the 'Added-Value' of a Right To Data Protection in the Eu Legal Order. *International and Comparative Law Quarterly*, *63*(03), 569–597. https://doi.org/10.1017/S0020589314000244

Lyrebird. (2017). *Lyrebird*. https://www.descript.com/lyrebird

Macionis, J. J., & Gerber, L. M. (1997). Sociology. Sociology, seventh Canadian edition. *Don Mills: Pearson Education Canada*, *1*.

Markoff, J. (2011). Computer wins on 'jeopardy!': trivial, it's not. *New York Times*, *16*. Retrieved 24 August 2021, fromhttps://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html

Mata, J., de Miguel, I., Durán, R. J., Merayo, N., Singh, S. K., Jukan, A., & Chamania, M. (2018). Artificial intelligence (AI) methods in optical networks: A comprehensive survey. *Optical Switching and Networking*. https://doi.org/10.1016/j.osn.2017.12.006

Matei, A. (2018). *New technology is forcing us to confront the ethics of bringing people back* *from the dead*. Retrieved 22 August 2021, from https://qz.com/896207/death-technology-will-allow-grieving-people-to-bring-back-their-loved-ones-from-the-dead-digitally/

Mayer-Schonberger, V., & Padova, Y. (2015). Regime Change: Enabling Big Data through Europe's New Data Protection Regulation. *Colum. Sci. & Tech. L. Rev.*, *17*, 315.

Mccarthy, J., Minsky, M., & Shannon, C. E. (2006). A proposal for the Dartmouth summer research project on artificial intelligence - August 31, 1955. *AI magazine*, *27*(4), 12–14.

McCartney, M. (2017). Margaret McCartney: Innovation without sufficient evidence is a disservice to all. *BMJ*, *358*, j3980.

Mcconnell, K. (2007). Creating People for Popular Consumption: Echoes of Pygmalion and "The Rape of the Lock" in Artificial Intelligence: AI. *Journal of Popular Culture*, *40*(4), 683–699.

Mcduff, D., Kaliouby, R. El, Senechal, T., Amr, M., Cohn, J. F., & Picard, R. (2013). *Affectiva-MIT Facial Expression Dataset (AM-FED): Naturalistic and Spontaneous Facial Expressions Collected In-the-Wild*. https://doi.org/10.1109/CVPRW.2013.130

McKendrick, J. (2019). Why AI Should Rightfully Mean Augmented Intelligence, Not Artificial Intelligence. *Forbes*. Retrieved 22 August 2021, from https://www.forbes.com/sites/joemckendrick/2019/06/29/why-ai-should-rightfully-mean-augmented-intelligence-not-artificial-intelligence/#7743b8bb1008

Mead, G. H. (1934). *Mind, self and society* (Vol. 111). Chicago University of Chicago Press.

Microsoft. (2018). *Maximising the AI Opportunity. How to harness the potential of AI effectively and ethically*. Retrieved 22 August 2021, from https://info.microsoft.com/UK-DIGTRNS-CNTNT-FY19-10Oct-26-MaximisingtheAIOpportunity-AID-731692-MGC0003240_01Registration-ForminBody.html

Microsoft. (2019). *Microsoft Ethical Principles*. Retrieved 22 August 2021, from https://www.microsoft.com/en-us/ai/our-approach-to-ai

Mill, J. S. (1887). *The positive philosophy of Auguste Comte*. New York: C. Blanchard.

Monitor Research Group. (2018). *Mon(ito)r Research Group*. Retrieved 22 August 2021, from https://moniotrlab.ccis.neu.edu/

Murgia, M., & Harlow, M. (2019). How top health websites are sharing sensitive data with advertisers. *Financial Times*. Retrieved 22 August 2021, from https://www.ft.com/content/0fbf4d8e-022b-11ea-be59-e49b2a136b8d

Myers, M. D., & Avison, D. (2002). *Qualitative research in information systems: a reader*.

Sage.

Myers, M. D., & Klein, H. K. (2011). A set of principles for conducting critical research in information systems. *MIS quarterly*, 17-36.

Narayanan, A. (2018). *Tutorial: 21 fairness definitions and their politics*. Retrieved 22 August 2021, from https://www.youtube.com/watch?v=jIXIuYdnyyk

Narayanan, A. (2019). How to recognize AI snake oil. *Princeton University*. Retrieved 22 August 2021, from https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf

Neff, G., & Nagy, P. (2016). Automation, Algorithms, and Politics| Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication*, *10*(0), 17.

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). Exploring the filter bubble: the effect of using recommender systems on content diversity. *Proceedings of the 23rd International Conference on World Wide Web*, 677–686.

Nilsson, N. J. (2009). *The quest for artificial intelligence*. Cambridge University Press.

Nilsson, P. (2018, February). How AI helps recruiters track jobseekers' emotions. *Financial Times*. Retrieved 22 August 2021, from https://www.ft.com/content/e2e85644-05be-11e8-9650-9c0ad2d7c5b5

Nimala, K., & Jebakumar, R. (2021). Sentiment topic emotion model on students feedback for educational benefits and practices. *Behaviour & Information Technology*, *40*(3), 311–319.

Norvig, P., & Russell, S. J. (2016). *Artificial intelligence : a modern approach* (P. Norvig author & E. Davis author (eds.); Third edit). Harlow, Essex, England : Pearson.

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

Oates, B. J. (2005). *Researching information systems and computing*. Sage.

Office for AI. (2021). *Ethics, Transparency and Accountability Framework for Automated Decision-Making*. Gov.UK. Retrieved 22 August 2021, from https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making/ethics-transparency-and-accountability-framework-for-automated-decision-making

Oliver, D. (2019). David Oliver: Lessons from the Babylon Health saga. *BMJ*, *365*, l2387.

Olson, P. (2018). This Health Startup Won Big Government Deals—But Inside, Doctors Flagged Problems. *Forbes*. Retrieved 22 August 2021, from https://www.forbes.com/sites/parmyolson/2018/12/17/this-health-startup-won-big-government-dealsbut-inside-doctors-flagged-problems/#2d71129ceabb

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). *WaveNet: A Generative Model for Raw Audio*. 1–15. *arXiv:1609.03499*.

Open Banking Working Group. (2018). *The Open Banking Standard*. working paper. Retrieved 22 August 2021, from https://www.scribd.com/doc/298569302/The-Open-Banking-Standard

Orbit. (2020). *Promoting RRI across the ICT Research Communit*. Retrieved 22 August 2021, from https://www.orbit-rri.org

Organisation for Economic Co-operation and Development (OECD). (2019). *Recommendation of the Council on Artificial Intelligence*. Retrieved 22 August 2021, from https://legalinstruments.oecd.org/api/print?ids=648&lang=en

Orlikowski, W. J., & Baroudi, J. J. (1991). Studying information technology in organizations: Research approaches and assumptions. *Information Systems Research*, *2*(1), 1–28. https://doi.org/10.1287/isre.2.1.1

Orlowski, A. (2016). *What's that, Adobe? A Photoshop for faking voices?* The Register. Retrieved 22 August 2021, from https://www.theregister.co.uk/2016/11/09/whats_that_adobe_a_photoshop_for_faking_voices/

Ovid, 43 B.C.-17 or 18 A D. (1968). *The Metamorphoses of Ovid* (W. Caxton 1422?-1491 (Ed.)). New York : George Braziller.

Owen, R. (2014). Responsible Research and Innovation: options for research and innovation policy in the EU. *European Research and Innovation Area Board (ERIAB), Foreword Visions on the European Research Area (VERA)*.

Packin, N. G. (2019). Algorithmic Decision-Making: The Death of Second Opinions? *New York University Journal of Legislation and Public Policy.*

Panch, T., Mattie, H., & Celi, L. A. (2019). The "inconvenient truth" about AI in healthcare. *NPJ Digital Medicine*, *2*(1), 1–3.

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, *45*(3), 438–450.

Parliament, T. H. E. E., Council, T. H. E., The, O. F., & Union European. (2003). *the European Parliament and the Council of the*. *2014*(April), 24–38. Retrieved 22 August 2021, from http://webarchive.nationalarchives.gov.uk/20121212135622/http://www.bis.gov.uk/files/file29931.pdf

Pasquale, F. (2015). *The black box society : the secret algorithms that control money and information*. Cambridge : Harvard University Press .

Podesta, J., Pritzker, P., Moniz, E. J., Holdren, J., & Zients, J. (2014). Big Data: Seizing Opportunities. *Executive Office of the President of USA*, *May*, 1–79. Retrieved 22 August 2021, from https://doi.org/10.5121/ijgca.2012.3203

Poole, D. L., Mackworth, A. K., & Goebel, R. (1998). *Computational intelligence: a logical approach* (Vol. 1). Oxford University Press New York.

Powles, J., & Hodson, H. (2017). Google DeepMind and healthcare in an age of algorithms. *Health and Technology*, *7*(4), 351–367. https://doi.org/10.1007/s12553-017-0179-1

Pozzebon, M. (2004). Conducting and evaluating critical interpretive research: Examining criteria as a key component in building a research tradition. In *Information systems research* (pp. 275-292). Springer, Boston, MA.

Privacy by Design Foundation. (2020). *About IRMA*. Retrieved 22 August 2021, from https://privacybydesign.foundation/irma-en

Privacy International. (2018). *PRESS RELEASE: Privacy International files complaints against seven companies for wide-scale and systematic infringements of data protection law*. Retrieved 22 August 2021, from https://privacyinternational.org/press-release/2424/press-release-privacy-international-files-complaints-against-seven-companies

Publication Office of the European Union. (2018). *25 years of the EU Single Market: Key Achievements*. Retrieved 22 August 2021, from http://www.europarl.europa.eu/resources/library/media/20180116RES91806/20180116RES91806.pdf

Ram, A. (2018). DeepMind develops AI to diagnose eye diseases. *Financial Times*. Retrieved 22 August 2021, from https://www.ft.com/content/84fcc16c-0787-11e8-9650-9c0ad2d7c5b5

Ramasubramanian, K., & Singh, A. (2017). *Machine learning using R*. Springer. New Delhi, India: Apress.

Raul, A. C. (2018). *The privacy, data protection and cybersecurity law review*. Law Business Research Limited.

Real Eyes. (2018). *Real Eyes*. Real Eyes Website. Retrieved 22 August 2021, from https://www.realeyesit.com/

Reed, C. (2018). How should we regulate artificial intelligence? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128), 20170360.

Reese, B. (2018). *A Conversation with Jakob Uszkoreit*. Voices in AI. Retrieved 22 August 2021, from https://gigaom.com/2018/10/04/voices-in-ai-episode-70-a-conversation-with-jakob-uszkoreit/

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018a). Algorithmic impact assessments: a practical framework for public agency accountability. *AI Now Institute*, 1-22.

Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018b). *Algorithmic Impact Assessments: a Practical Framework for Public Agency Accountability*. *April*. Retrieved 22 August 2021, from https://ainowinstitute.org/aiareport2018.pdf

Rezende, I. N. (2020). Facial recognition in police hands: Assessing the 'Clearview case'from a European perspective. *New Journal of European Criminal Law*, *11*(3), 375–389.

Rhoen, M. (2016). Beyond consent: improving data protection through consumer protection law. *Journal on Internet Regulation - Leiden Law School*, *5*(1), 1–15. https://doi.org/10.14763/2016.1.404

Rich, E., & Knight, K. (1991). Artificial intelligence. McGraw-Hill.

Richardson, H., & Robinson, B. (2007). The mysterious case of the missing paradigm: a review of critical information systems research 1991–2001. *Information Systems Journal*, 17(3), 251-270.

Ridder, H.-G. (2017). The theory contribution of case study research designs. *Business Research*, *10*(2), 281–305.

Robotics Openletter EU. (2018). *OPEN LETTER TO THE EUROPEAN COMMISSION ARTIFICIAL INTELLIGENCE AND ROBOTICS*. Retrieved 26 August 2021, from https://archive.ph/cycYW (originally http://www.robotics-openletter.eu/)

Robson, C. (2011). *Real world research: A resource for users of social research methods in applied settings 3rd edition*. West Sussex: John Wiley & Sons.

Romele, A. (2019). *Digital Hermeneutics: Philosophical Investigations in New Media and Technologies*. Routledge

Roosendaal, A. (2018). *DPIA Diagnostic Data in Microsoft Office Proplus Commissioned by the Ministry of Justice and Security for the benefit of SLM Rijk Vendor Management (Strategic Microsoft Dutch Government) Sjoera Nas*. *November*. Retrieved 22 August

2021, from https://www.rijksoverheid.nl/binaries/rijksoverheid/documenten/rapporten/2018/11/07/data-protection-impact-assessment-op-microsoft-office/DPIA+Microsoft+Office+2016+and+365+-+20191105.pdf

Rouvroy, A., & Poullet, Y. (2009). The right to informational self-determination and the value of self-development: Reassessing the importance of privacy for democracy. In *Reinventing data protection?* (pp. 45–76). Springer, Dordrecht.

Rowe, F., EL AMRANI, R., Limayem, M., Dennis, A., & Introna, L. (2020). *ECIS2020 Welcome and Keynote Address*. Retrieved 22 August 2021, from https://aisel.aisnet.org/ecis2020_sessionrecordings/3/

RRI Tools. (2020). *RRI FOR BUSINESS & INDUSTRY*. Retrieved 22 August 2021, from https://www.rri-tools.eu/da_DK/business-and-industry

Ruce, P. J. (2011). Anti-money laundering: The challenges of know your customer legislation for private bankers and the hidden benefits for relationship management (the bright side of knowing your customer). *Banking LJ*, *128*, 548.

Ryan, M., & Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, *786641*. https://doi.org/10.1108/JICES-12-2019-0138

Ryan, J. (2018). *Regulatory complaint concerning massive, web-wide data breach by Google and other "ad tech" companies under Europe's GDPR*. Retrieved 22 August 2021, from https://brave.com/adtech-data-breach-complaint/

Sample, I., & Hern, A. (2014). *Scientists dispute whether computer "Eugene Goostman" passed Turing test*. Retrieved 22 August 2021, from https://www.theguardian.com/technology/2014/jun/09/scientists-disagree-over-whether-turing-test-has-been-passed

Samuel, a L. (1959). Some Studies in Machine Learning. *IBM Journal of research and development*, *3*(3), 210-229.

Sánchez-Monedero, J., & Dencik, L. (2020). The politics of deceptive borders:'biomarkers of deceit'and the case of iBorderCtrl. *Information, Communication & Society*, 1–18.

Saunders, M., Lewis, P., & Thornhill, A. (2015). *Research methods for business students*. Pearson education.

Sayburn, A. (2017). Will the machines take over surgery? *The Bulletin of the Royal College of Surgeons of England*, *99*(3), 88–90. https://doi.org/10.1308/rcsbull.2017.87

Schmerheim, P. (2018). The Palgrave handbook of posthumanism in film and television. In *New Review of Film and Television Studies* (Vol. 16, Issue 1, pp. 89–92). Routledge. https://doi.org/10.1080/17400309.2018.1426155

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Seargeant, P., & Tagg, C. (2019). Social media and the future of open debate: A user-oriented approach to Facebook's filter bubble conundrum. *Discourse, Context & Media*, *27*, 41–48.

Sejnowski, T. J. (2018). *The deep learning revolution*. Mit Press.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68.

Selbst, A., & Powles, J. (2017). Meaningful Information and the Right to Explanation. *International Data Privacy Law*, *7*(4), 233. https://doi.org/10.1093/IDPL/IPX022

Shead, S. (2017). *Google DeepMind is giving the NHS free access to its patient monitoring app*. Business Insider UK. Retrieved 22 August 2021, from https://www.businessinsider.com/google-deepmind-is-giving-three-nhs-trusts-free-access-to-its-streams-app-2017-6?r=US&IR=T

Shearing, H. (2019). *Millions Of Students' Sexual Orientations And Religious Beliefs Are Being Held On A Government Database*. Buzzfeed. Retrieved 22 August 2021, from https://www.buzzfeed.com/hazelshearing/the-government-has-a-database-of-millions-of-students

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., Van Den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359.

Sim, K., Brown, A., & Hassoun, A. (2021). Thinking Through and Writing About Research Ethics Beyond" Broader Impact". *ArXiv Preprint ArXiv:2104.08205*.

Skirpan, M., & Gorelick, M. (2017). *The Authority of "Fair" in Machine Learning*. http://arxiv.org/abs/1706.09976.

Smallman, M., Lomme, K., & Faullimmel, N. (2015). Report on the analysis of opportunities, obstacles and needs of the stakeholder groups in RRI practices in Europe. *RRI Tools—Fostering Responsible Research and Innovation. University College London.* Retrieved 22 August 2021, from https://rri-tools.eu/documents/10184/193151/2_RRITOOLS-wp2-

D_2.2.pdf/f258550d-ba1f-43b2-bec2-f7e1152aa79c

Smith, R. (1989). *Prior Analytics*. Hackett Publishing.

Sobel, B. (2020). A New Common Law of Web Scraping. Retrieved 23 August 2021, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3581844

Sofka, C. J., Gibson, A., & Silberman, D. R. (2017). Digital immortality or digital death?: Contemplating digital end-of-life planning. *Postmortal Society: Towards a Sociology of Immortality*, *2017*, 173–196. https://doi.org/10.4324/9781315601700

Somerville, H. (2018). *California proposes new rules for self-driving cars to pick up passengers*. Reuters. Retrieved 22 August 2021, from https://www.reuters.com/article/us-selfdriving-regulations-california/california-proposes-new-rules-for-self-driving-cars-to-pick-up-passengers-idUSKCN1HD2ZL

Stahl, B. (2018). RRI in Industry. *ORBIT Journal*, *1*(3). https://doi.org/10.29297/orbit.v1i3.64

Stahl, B. C. (2008). The ethical nature of critical research in information systems. *Information Systems Journal*, *18*(2), 137–163.

Stahl, B. C. (2012). Responsible research and innovation in information systems. *European Journal of Information Systems,* 21:3, 207-211, DOI: 10.1057/ejis.2012.19

Stahl, B. C. (2013). Responsible research and innovation: The role of privacy in an emerging framework. *Science and Public Policy*, *40*(6), 708–716.

Stahl, B. C. (2021a). *EU is cracking down on AI, but leaves a loophole for mass surveillance – Bernd Carsten Stahl*. Retrieved 22 August 2021, from https://inforrm.org/2021/04/22/eu-is-cracking-down-on-ai-but-leaves-a-loophole-for-mass-surveillance-bernd-carsten-stahl/

Stahl, B. C. (2021b). *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies* (p. 124). Springer Nature.

Stahl, B. C., Obach, M., Yaghmaei, E., Ikonen, V., Chatfield, K., & Brem, A. (2017). The responsible research and innovation (RRI) maturity model: Linking theory and practice. *Sustainability (Switzerland)*, *9*(6). https://doi.org/10.3390/su9061036

Stahl, B. C., & Wright, D. (2018a). Ethics and privacy in AI and big data: Implementing responsible research and innovation. *IEEE Security & Privacy*, *16*(3), 26–33.

Stahl, B. C., & Wright, D. (2018b). *Proactive Engagement with Ethics and Privacy in AI and Big Data-Implementing responsible research and innovation in AI-related projects*.

Stake, R. E. (2013). *Multiple case study analysis*. Guilford Press.

Stalla-Bourdillon, S., Pearce, H., & Tsakalakis, N. (2018). The GDPR: A game changer for electronic identification schemes? The case study of Gov. UK Verify. *Computer Law & Security Review*, *34*(4), 784–805.

Sticky. (2017). *Sticky*. StickyWebsite. Retrieved 22 August 2021, from https://www.sticky.ai/

Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum*, *56*(4), 24–31.

Strübing, J. (1998). Bridging the Gap: On the Collaboration between Symbolic Interactionism and Distributed Artificial Intelligence in the Field of Multi-Agent Systems Research. *Symbolic Interaction*, *21*(4), 441–463.

Su, Y.-S., Suen, H.-Y., & Hung, K.-E. (2021). Predicting behavioral competencies automatically from facial expressions in real-time video-recorded interviews. *Journal of Real-Time Image Processing*, 1–11.

Tegmark, M. (2017). *Life 3.0 : being human in the age of artificial intelligence*. London : Allen Lane.

The Boston Consulting Group. (2018). *The Most Innovative Companies 2018: Innovators Go All In On Digital*. Retrieved 22 August 2021, from http://image-src.bcg.com/Images/BCG-Most-Innovative-Companies-Jan-2018_tcm9-180700.pdf

The Economist. (2017). Data is giving rise to a new economy. *The Economist.* Retrieved 22 August 2021, from https://www.economist.com/briefing/2017/05/06/data-is-giving-rise-to-a-new-economy

The Economist. (2018a). Big tech is growing, but so is investors' caution. *The Economist*. Retrieved 22 August 2021, from https://www.economist.com/news/business/21741189-years-american-tech-giants-were-treated-single-asset-class-no-more-big-tech

The Economist. (2018b, March). Non-tech businesses are beginning to use artificial intelligence at scale. *The Economist*. Retrieved 22 August 2021, from https://www.economist.com/news/special-report/21739431-artificial-intelligence-spreading-beyond-technology-sector-big-consequences

The Economist Science and Technology. (2018). For artificial intelligence to thrive, it must explain itself. *The Economist*. Retrieved 22 August 2021, from https://www.economist.com/news/science-and-technology/21737018-if-it-cannot-who-will-trust-it-artificial-intelligence-thrive-it-must

Regulation EU 910/2014 of the European Parliament and of the Council of 23 July 2014 on electronic identification and trust services for electronic transactions in the internal market and repealing Directive 1999/93/EC (eIDAS Regulation), European Union 44 (2014). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3AOJ.L_.2014.257.01.0073.01.ENG

Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, (2016) https://eur-lex.europa.eu/eli/reg/2016/679/oj

The Financial Industry Regulatory Authority. (2011). *Know Your Customer and Suitability*. Retrieved 22 August 2021, from https://www.finra.org/sites/default/files/NoticeDocument/p122778.pdf

The Human Brain Project. (2021). *Welcome to the Human Brain Project*. Retrieved 22 August 2021, from https://www.humanbrainproject.eu/en/

The Royal Society. (2017). *Machine learning : the power and promise of computers that learn by example*. Retrieved 22 August 2021, from https://royalsociety.org/-/media/policy/projects/machine-learning/publications/machine-learning-report.pdf

The Electronic Identification and Trust Services for Electronic Transactions Regulations 2016, (2016) (testimony of The UK Parliament). Retrieved 22 August 2021, from https://www.legislation.gov.uk/uksi/2016/696/pdfs/uksi_20160696_en.pdf

The UK Parliament. (2017). *Professor Lilian Edwards – Written evidence (AIC0161)*. Retrieved 22 August 2021, from http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69669.html

The UK Parliament AI Committee. (2017a). *DeepMind – Written evidence (AIC0234)*. Retrieved 22 August 2021, from fromhttps://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwiMzZvA_vbbAhUMDMAKHZZpCbIQFggpMAA&url=https%3A%2F%2Fpublications.parliament.uk%2Fpa%2Fld201719%2Fldselect%2Fldai%2F100%2F10011.htm&usg=AOvVaw209lws59MB-JwJXWIES2HL

The UK Parliament AI Committee. (2017b). *IEEE European Public Policy Initiative Working Group on ICT – Written evidence (AIC0106)*. Retrieved 23 August 2021, from https://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69590.html

Tidy, J. (2020). *Blackbaud hack: More UK universities confirm breach*. BBC News. Retrieved 23 August 2021, from https://www.bbc.co.uk/news/technology-53528329

Trade Union Congress (TUC). (2020). *Technology managing people. The worker experience*. Retrieved 23 August 2021, from https://www.tuc.org.uk/sites/default/files/2020-11/Technology_Managing_People_Report_2020_AW_Optimised.pdf

Trueimpact. (2017). *Trueimpact*. Trueimpact Website. Retrieved 23 August 2021, from http://www.trueimpact.ca/

Tsakalakis, N., O'hara, K., & Stalla-Bourdillon, S. (2016). Identity Assurance in the UK: technical implementation and legal implications under the eIDAS Regulation. In *Proceedings of the 8th ACM Conference on Web Science*, 55–65.

Tufekci, Z. (2017). *We're building a dystopia just to make people click on ads*. Ted Talk. Retrieved 23 August 2021, from https://www.ted.com/talks/zeynep_tufekci_we_re_building_a_dystopia_just_to_make_people_click_on_ads/up-next

Tufekci, Z. (2018). Facebook's Surveillance Machine. *The New York Times*. Retrieved 23 August 2021, from https://www.nytimes.com/2018/03/19/opinion/facebook-cambridge-analytica.html?rref=collection%2Fcolumn%2Fzeynep-tufekci&action=click&contentCollection=opinion&region=stream&module=stream_unit&version=latest&contentPlacement=1&pgtype=collection

Turing, A. (1950). Turing. Computing machinery and intelligence. *Mind*, *59*(236), 433–460. https://doi.org/http://dx.doi.org/10.1007/978-1-4020-6710-5_3

Turner, J. (2018). *Robot rules: Regulating artificial intelligence*. Springer.

Tutt, A. (2017). An FDA for algorithms. *Admin. L. Rev.*, *69*, 83.

Tynan, D. (2016). Augmented eternity: scientists aim to let us speak from beyond the grave. *The Guardian*. Retrieved 23 August 2021, from https://www.theguardian.com/technology/2016/jun/23/artificial-intelligence-digital-immortality-mit-ryerson

Uebel, T. (2006). *Vienna circle*. The Stanford Encyclopedia of Philosophy Fall 2006 Edition, ed. Zalta, E. N.. Retrieved 23 August 2021, from http://plato.stanford.edu/archives/fall2006/entries/vienna-circle

UK Department for Digital, Culture, Media & Sport. (2018). *Data Ethics Workbook*. Retrieved 23 August 2021, from https://www.gov.uk/government/publications/data-ethics-workbook

UK Department for Digital, Culture, Media & Sport. (2020). *THE DATA PROTECTION ACT 2018 KEELING SCHEDULE*. Retrieved 21 August 2021, from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/969513/20201102_-_DPA_-__MASTER__Keeling_Schedule__with_changes_highlighted__V4.pdf

UK Department for Digital, Culture, Media & Sport. (2021). *Digital Regulation: Driving* *growth and unlocking innovation*. Retrieved 23 August 2021, from https://www.gov.uk/government/publications/digital-regulation-driving-growth-and-unlocking-innovation

United Nation General Assembly (1948). *Universal Declaration of Human Rights.* UN General Assembly, 302(2), 14-25.

University of Oxford - Future of Humanity Institute. (2019). *Future of Humanity Institute*. Retrieved 23 August 2021, from https://www.fhi.ox.ac.uk/

Unruly. (2018). *Unruly*. Unruly Website. Retrieved 23 August 2021, from https://unruly.co/

Vaismoradi, M., Jones, J., Turunen, H., & Snelgrove, S. (2016). *Theme development in qualitative content analysis and thematic analysis*. J. Nurs. Educ. Pract. 6, 100–110.

van den Hoven, J. (2013). *Options for strengthening responsible research and innovation: report of the Expert Group on the State of Art in Europe on Responsible Research and Innovation*. Publications Office of the European Union. Retrieved 23 August 2021, from https://op.europa.eu/en/publication-detail/-/publication/1e6ada76-a9f7-48f0-aa86-4fb9b16dd10c

Van Kleek, M., Seymour, W., Veale, M., Binns, R., & Shadbolt, N. (2018). The Need for Sensemaking in Networked Privacy and Algorithmic Responsibility. *Sensemaking in a Senseless World: Workshop at ACM CHI'18, 22 April 2018, Montréal, Canada*. Association for Computing Machinery (ACM).

Veale, M., Binns, R., & Ausloos, J. (2018). When Data Protection by Design and Data Subject Rights Clash. *International Data Privacy Law*. *8*(2), 105-123. https://doi.org/10.1093/idpl/ipy002/4960902

Veale, M., Binns, R., & Van Kleek, M. (2018). *Some HCI Priorities for GDPR-Compliant Machine Learning. arXiv preprint arXiv:1803.06174*

Veale, M., & Edwards, L. (2018). Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Computer Law and Security Review*, *34*(2), 398–404. https://doi.org/10.1016/j.clsr.2017.12.002

Verganti, R., Vendraminelli, L., & Iansiti, M. (2020). Innovation and design in the age of artificial intelligence. *Journal of Product Innovation Management*, *37*(3), 212–227.

Wachter, S., & Mittelstadt, B. (2019). A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Colum. Bus. L. Rev.*, 494.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76–99.

Warwick, K., & Shah, H. (2016). Can machines think? A report on Turing test experiments at the Royal Society. *Journal of Experimental & Theoretical Artificial Intelligence*, *28*(6), 989–1007.

Waters, R. (2017). How machine learning creates new professions — and problems. *The Financial Times*. Retrieved 23 August 2021, from https://www.ft.com/content/49e81ebe-cbc3-11e7-8536-d321d0d897a3

Waters, R. (2018, April 22). Google under fire over tactics for EU data regulation. *Financial Times*. Retrieved 23 August 2021, from https://www.ft.com/content/898c812a-4640-11e8-8ee8-cae73aab7ccb

Watson, G. J., Desouza, K. C., Ribiere, V. M., & Lindič, J. (2021). Will AI ever sit at the C-suite table? The future of senior leadership. *Business Horizons*. *Business Horizons*, *64*(4), 465-474.

Webb, H. M., Koene, A., Patel, M., & Vallejos, E. P. (2018). Work-in-progress: Multi-stakeholder dialogue for policy recommendations on algorithmic fairness. In *Proceedings of the 9th International Conference on Social Media and Society (SMSociety '18). Association for Computing Machinery,* New York, NY, USA, 395–399. DOI:https://doi.org/10.1145/3217804.3217952

Webber, M. (2016). The GDPR's impact on the cloud service provider as a processor. *Privacy & Data Protection*, *16*(4), 1–5.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation.* W. H. Freeman & Co.

Weizenbaum, J. (1978). Once more—a computer revolution. *Bulletin of the Atomic Scientists*, *34*(7), 12–19. https://doi.org/10.1080/00963402.1978.11458531

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., ... & Schwartz, O. (2018). *AI now report 2018* (pp. 1-62). New York: AI Now Institute at New York University. Retrieved 26 August 2021, from https://ainowinstitute.org/AI_Now_2018_Report.pdf

Williams, C. (2007). Research methods. *Journal of Business & Economic Research*, *5*(3), 65–72.

Wilson, H. J., & Daugherty, P. R. (2018). Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*. Retrieved 22 August 2021, from https://hbr.org/2018/07/collaborative-intelligence-humans-and-ai-are-joining-forces

Winston, P. (1992). Learning by building identification trees. *Artificial Intelligence*, 423–442.

Xu, D., Nair, S., Zhu, Y., Gao, J., Garg, A., Fei-Fei, L., & Savarese, S. (2018). Neural Task Programming: Learning to Generalize Across Hierarchical Tasks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3795-3802). http://arxiv.org/abs/1710.01813

Yin, R. K. (2003). Case study research design and methods third edition. *Applied Social Research Methods Series*, *5*.

Yin, R. K. (2014). *Case study research : design and methods* (Fifth edit). Los Angeles : SAGE.

Yin, R. K. (2018). *Case study research and applications : design and methods* (Sixth edit). Los Angeles : SAGE.

Zagorecki, A. T., Johnson, D. E. ., & Ristvej, J. (2013). Data mining and machine learning in the context of disaster and crisis management. *International Journal of Emergency Management*, *9*(4), 351. https://doi.org/10.1504/IJEM.2013.059879

Zhang, D., Mishra, S., Brynjolfsson, E., Etchemendy, J., Ganguli, D., Grosz, B., Lyons, T., Manyika, J., Niebles, J. C., & Sellitto, M. (2021). The AI Index 2021 Annual Report. ArXiv Preprint ArXiv:2103.06312.

Zheng, Y., & Stahl, B. C. (2011). Technology, capabilities and critical perspectives: what can critical theory contribute to Sen's capability approach?. *Ethics and Information Technology*, 13(2), 69-80.

Zhu, J., & Harrell, D. F. (2009). The artificial intelligence (ai) hermeneutic network: A new approach to analysis and design of intentional systems. *Proceedings of the 2009 Digital Humanities Conference*, 301–304.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.

APPENDICES

Appendix A: Papers, conferences, workshops

Papers

- Addis, C., & Kutar, M. S. (2018, March). The general data protection regulation (GDPR), emerging technologies and UK organisations: awareness, implementation and readiness. In *UK Academy for Information Systems Conference Proceedings* 2018 (p. 29). UKAIS–UK Academy for Information Systems.
- Addis, C., & Kutar, M. (2019, August). AI management an exploratory survey of the influence of GDPR and FAT principles. In 2019 *IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (pp. 342-347). IEEE.
- Addis, C., & Kutar, M. S. (2020, April). General Data Protection Regulation (GDPR), Artificial Intelligence (AI) and UK organisations: a year of implementation of GDPR. In *UK Academy for Information Systems Conference Proceedings 2020* (p. 24). UKAIS–UK Academy for Information Systems.

Conferences

- The UK Academy for Information Systems/UKAIS Conference (Oxford, 2020).
- IEEE Smart World Congress Forum on Ethics and Human Rights in Smart Information Systems (Leicester, 2019).
- Interdisciplinary Summer School Governance Technologies: Privacy, Fairness & Transparency (Hamburg, Germany, Sep 2019).
- Salford Postgraduate Annual Researcher Conference (SPARC) (University of Salford, 2019).
- The UK Academy for Information Systems/UKAIS Conference (Oxford, 2018).

Organisation of workshop

- GDPR: Exploring practical impact and implementation (University of Salford - 2017-2018). Event created in collaboration with Dr Maria Kutar (Supervisor).

Reviews of academic articles

- International Conference on Information Systems/ISIS (2020-21).
- Journal of Cyber Policy (2020).

Appendix B: Case studies - documents and observations

1. Case Study 1 (CS1)

1.1 Documents

In the following sections, a detailed analysis of the key documents is presented. These include Learner Analytics and Attendance Monitoring business cases, DPIA, user guide and privacy policy.

a. Business case. Attendance Monitoring Project (BC-AM)

The Attendance Monitoring project and the LA project are formally two different projects, but they share the same Project Board and are strictly connected.

This project aims to improve the retention of students, and this is strictly connected to their engagement. Their attendance is considered a "*key metric*" (BA-AM p1) of their engagement, and it is measured by "*increasing the quality and accessibility of information*" (Ibidem), which is then used to support early interventions. By recording the participation of students, the organisation plans to: identify students at risk; arrange appropriate interventions; inform students about their attendance records; enhance their retention; provide information on academic success and student employability.

Transparency is a key element, as higher visibility and access to attendance data (by organisation and students) is considered critical. The possibility to opt-out is not considered. While an exception solution for students without a mobile is said to be identified, the same is not planned for those who do not want to provide their personal data.

The project is also significant in terms of its importance for business improvements, brand reputation and economic benefits associated with increased retention. The accuracy of data is also envisaged as being capable of increasing organisational efficiency, optimising resources, and reducing costs. The project aims at replicating existing processes and releasing staff capacity to facilitate student-facing intervention.

b. Business case - LA Project (BC-LA)

The second project aims at *"enabling earlier intervention by identifying at risk students before risks become an issue"* (BC-LA, p3). This is done by providing evidence, predicting risks, responding fast, and preventing issues.

Data gathered from existing data feeds is used to train the system for identifying at risk students and to improve their outcomes. LA aims at getting "*the right data to the right people*" (Ibidem, p2) in order to "*enable evidence-based strategic decisions*"(Ibidem). Reports and visualisations will be tested "*throughout the agile process to ensure a good end-user experience*" (Ibidem). The project is said to provide a quick response and increase the organisation's reputation for responsiveness. It is also seen as being able to increase compliance and demonstrate a "new culture of quick delivery". A coordinated and coherent approach to strategic data reporting is ensured by the project Steering Board and project team which cover all key areas of the organisation. An embedded system security is said to be realised via technical engagement on security, and via an engagement with Information Governance from the start of the project. An important collaboration also being considered is looking at supporting the *"ability to add further data sources"* (Ibidem, p3).

The project will follow the same governance practices and controls as the programme it is part of (periodical reports, dashboard updates, programme risk and issue log, joint dependency management, monthly Project Steering Board meetings). Training is assumed to be delivered internally at no cost. No specific training upskilling is envisaged. There are only technical risks. The document mentions collection of special categories of data, future increment of data, and third parties which will be managing personal data (full IT support function provided by a mix of in-house and third parties). However, no specific risks related to ML or DP are envisaged.

Like the Attendance monitoring project, this app is a cloud-based solution (MLaaS).

c. DPIA

The DPIA for engagement analytics was created by one senior manager. The document appears to be a standard form used for different types of projects and it includes two parts: general guidance and the assessment of the project.

A. General guidance. The first part of the document lists some general rules for performing DPIAs created by the organisation. This part reveals some interesting details about the importance of DPIAs for the organisation.

When? The DPIA should be created early in the project, run alongside the development process, initiated prior to processing data, and updated every time the requirements become clearer. The document includes a link to an ICO page on privacy by design. No specific link to DPIA guidance is provided.

Who and how? The Project Owner creates and updates the DPIA with the support of the Project Manager, as they know about the project scope, background, system, data, and information flows. The document is said to a be collaborative document. Digital IT is mentioned as a key stakeholder, Information Governance "*retain a copy*" and the DPO signs it off. DP roles are not mentioned as active stakeholders.

External entities: As the involvement of external subjects increases the risk, the creation of a detailed DPIA and a contract are advised. There is no mention of specific documents, such as DPIA, required from partners.

Obligation: The document includes a reminder that failing to perform a DPIA can impact delivery and future compliance. The document is considered a necessary milestone before moving to the next stage.

Infographic: An infographic is provided to help to identify the areas of risk: data minimisation, purpose, information for data subjects, audit, access, data location, retention and some data subject rights. These are standard areas of risk used for any type of project. No risks linked to specific technologies are considered.

B. Impact Assessment of the project

Releases: LA is said to create a cloud based end-to-end analytics solution using ML technology. Three releases are planned:

1st: Creation of the new common data layer from current different systems.

2nd.: Integration with a new Attendance Monitoring Azure SQL Database (chosen for its sustainability and scalability). This will include the student location data received from the app of a third party (GPS-based Mobile App) that collects data for the organisation.

3rd: Creation of ML functionality, "*likely to include sourcing incremental personal/sensitive socio-demographic data about the students to inform analytical profiling"* (DPIA, p5).

Processed Data: *Input data*. Different types of personal data are fed into the system:

- Student ID, University E-Mail, Personal E-Mail, Mobile Phone Number, Date of Birth, Home Address, Tier 4 Visa * (special category of data), Statement of Access * (special category of data), Term, Degree, School, Study Year, Study Mode, Study Level, Registration Status, Registration Date, Modules, Submission timelines (on time or late), Submission scores (pre-moderated), and Plagiarism detection service.
- The number of minutes spent on Wi-Fi (inside and outside campus) and reviewing course and video content in the VLE.
- The number of print requests, physical withdrawals from the Library on Campus, and the number of digital logins to Library materials.

Therefore, a high level of granular information is collected. This creates further risk related to monitoring due to its intensification. This can create another issue linked to the potential possibility that students will disagree with the increased monitoring. The risk of this is considered to be very low, as students are said to be reasonably expecting such data processing to take place because they are informed about it when they enrol. The risk is mitigated by reassuring them about the good intentions of the organisation.

Data Quality - The application is aggregating data directly from the source systems, and it is inheriting any additions, deletions, or amendments to the data. Potential data errors can create an issue for the reporting system. This is considered to be a low risk, being mitigated by DIT support teams.

No indication is given about the kind of control DIT is called to perform. No other risks are identified in relation to the creation of the common data layer.

Output data - The application calculates the attainment score, which is evaluated and acted upon by the school administrators. The score is used in attendance, engagement, and intervention reporting. No risks are identified in relation to data aggregation or the calculation of the engagement score.

Benefits: The key benefits for the organisation are the minimisation of loss of revenue and the creation of a "*skillset and capability*"(DPIA, p6) that can be re-used for other projects.

The benefits for the students are increased welfare, satisfaction and advocacy of the University.

Main stakeholders: The key actors are the Business Owner, Project Manager, Project Sponsor, CIO, and delivery partners who are creating the ML capability. Information Governance, DPO, Users and other third parties are not included.

The involvement of the DPO is only requested in relation to data retention and erasure, critical risks which need the DPO's guidance.

Third parties: The organisations handling personal data are G-Cloud 9/Azure and GPS-based Mobile App.

1. G-Cloud 9: DIT is in charge of conducting standard due diligence (when initially engaging), and the management is in charge of controlling the adherence to the contract (attached to DPIA but not viewed by the researcher). The identified risk is low.
2. No reference to the agreements with the GPS-based App company and no reference to the third party building the ML capacity.

GDPR: One of the assessment questions refers to the inclusion of GDPR elements into the procurement process. The question is not answered in the form. privacy by design is said to be considered within the project. privacy by default is not mentioned. The only mentioned lawful basis is contract. The risk is said to be low.

Risks: According to the process, risks are supposed to be identified and mitigated by the Business Owner, after discussing the DPIA with Digital IT and Information Governance.

Therefore, the process advocates a more collaborative process between different parts of the business. The only identified risk related to the system is a potential data breach from unauthorised access. It is medium and mitigated by role-based access.

No risk linked to the centralisation of personal data (i.e., hacking) is expected.

Project Stage: The project is set to be at the delivery stage.

Data storage: Data is stored in an EU server (Cloud, Azure).

Training: Staff training is planned to take place via post-implementation workshops, and the provision of a user guide. There is no mention of future specific workshops on AI/ML.

Data Access/SAR: School managers are expected to act after receiving the request via Information Governance.

Approval: The document was conditionally approved by the Information Governance Team, which requested the following:

- The retention rules are to be defined and applied before full rollout.
- The user guide is to specify the exact location of reports (only allowed on server on campus/Office 365, and not on the external cloud).
- Further revisions are planned, especially the third that will include ML.

No other comments were made on the document, nor on the need to involve the DPO to develop other parts of the assessment.

d. User guide

The definition of the cohort includes all students on the same course, study year and term.

Some problems with this definition are already acknowledged in the guide.

- It is not ideal for postgraduate programmes and for any other courses with a very low number of students (as a small cohort can affect the accuracy of the data).
- No references to disabilities, or how different disabilities could impact on the measurement of attendance (i.e., vision impairments and printing) are found.
- DP measures focused on regulating data access to data and identifying the most common risks of breaches.
- The time limit for reporting data breaches is missing.

It is unknown whether Information Governance or the DPO had provided any input on the creation of the guide.

e. Student privacy policy (SPP)

Some parts of the policy are relevant for the project:

Legal bases. Processing students' data is justified by two legal bases: public task and contract.

The current lawful bases can justify the processing of engagement data. However, the processing of student data for prediction performed by the ML system is likely to require further modifications of the privacy policy.

Data. Amongst the special categories of data collected are the political beliefs of students, the specific purpose of which is left unclear.

Purposes. Amongst the purposes included are engagement monitoring, student wellbeing and support. Data is also used to "*build a profile of commonly shared characteristics for marketing purposes"* (SPP).

Retention period. Data is retained for as long as it is required to perform its purpose or for as long as is required by law (student data is generally retained for six years). At the end of the retention period, the data will either be completely deleted or anonymised.

Third parties. It is ensured that students' data collected from third parties is lawfully provided.

Data is shared with third parties only to meet contractual needs or to improve the services.

Third parties *may* (SPP) use the data for the exact purposes specified in the contract signed with the parties. Data is either deleted or *anonymised* at the end of the contract. Data is also shared with parties "*with an interest in tracking student progress and attendance*" (SPP), for example, student and research sponsors, research councils, and the NHS, current or potential employers, to provide references or, for sponsored or placement students, to confirm details of progress and attendance.

1.2 Observations

The interaction with the participants was very positive. They were willing to participate in the research and to talk about their contribution, the project, and its importance for the organisation. All interviews were carried out face to face in different buildings belonging to the organisations, with only one carried out via Skype. The requests for the interviews were all sent by email. Most of the participants responded quickly, accepting the request. This was especially the case with the permanent staff of the organisation. Some respondents were not expecting to be involved in the research. This was mainly due to the perception of their roles as more marginal or limited to a specific task or phase.

However, as the questions aimed at looking at the involvement in the project but also at exploring the understanding of AI and DP, their inputs, recollections, and thinking were important. Everybody expressed the firm belief that the project was being developed for the purpose of improving the students' experience, and all showed enthusiasm for supporting them more efficiently.

Two interviews were carried out with the two leaders of the organisation. Their offices are located in a different building from other staff. The interviews were very different in terms of approach to the technology and its use for strategic and organisational development.

Although working very closely, the two participants had a very different understanding and perception of the capacities and the strategic role of ML in general, and in relation to the organisation. One leader could already envisage the impact on students, some future strategic use of the data, and the potential for the organisation while the other had a more cautious approach to ML, expressed via a long discussion on fairness and the risks of predicting future outcomes. The same leader could see how, paradoxically, the ultimate proof of successful use of ML would be if the organisation were capable of achieving a positive outcome, opposite to an original negative prediction on a student. The different kinds of discussion and depth of the two interviews meant that they were different in length.

Similarly, the interviews with the DPO and the Information Governance officer, both in charge of DP, were very different. They were located in different buildings, and they meet weekly to discusses the different cases related to DP and Freedom of Information. Their physical distance appeared to mirror the distance in their different approaches to DP with the first being more traditional and the second being more oriented towards emerging technologies.

The DPO appeared to be more focused on discussing DP in terms of security, internal data breaches and company culture while the Information Governance officer, who was new to the organisation, had a more layered approach to it, connecting DP to new technologies and identifying the impact on inequalities and some consequences in different sectors.

Additionally, two interviews were conducted with participants who had more technical roles. One of the interviews was conducted face to face, and the second was on Skype. Both were consultants temporarily employed to create specific elements of the system. They were building the technical capacities for the organisation, and they had long experience gained whilst working for various organisations. They were also less physically present on the premises, as they could do part of the work from home. Amongst the participants, they were the ones who had a more transient relationship to the project and the organisation. Thanks to their different relationship to the organisation and their experience with other companies, they were able to talk about the project, about their experiences on other projects and could make some comparisons between the different experiences. Their being insiders/outsiders to the project/organisation allowed them to view the project and the organisational internal dynamics also as observers. These two participants shared some common elements and some clear differences. One had a deep knowledge of ML and seemed less familiar with ML integration in organisations. The other was more knowledgeable of processes and less aware of AI/ML specificities. They both had a utilitarian approach to technology, which was considered to be neutral. They were focused on the data and system, and they appeared to be generally less concerned about biases.

And yet, one of the two, only at the end of one interview, warned about the risk of biases getting rapidly stronger in autonomous decision-making systems, but adding that '*probably'* other ML algorithms were being designed to avoid that scenario. This was again considered only a technical problem easily solvable with some data preparation.

A strong reaction occurred when one of the respondents was asked some questions about the GDPR. Due to the high sanctions brought up by the Regulation, they did not feel entitled to give any opinion, as they were used to doing prior GDPR, because the potential risk to the organisation was perceived now too high.

Also of interest was the interview with the person in charge of data governance, training, and development. They had gained experience in working for similar organisations and they were managing a big team. They had business and technical expertise, and their competence transpired while talking about the project. They were aware of diversity issues and specific challenges within the sector. They were in charge of the DPIA. They had created a few DPIAs for other projects and displayed a high level of confidence in doing them.

The interview with a manager who worked in the E&D area was surprisingly rich in information. They were relatively new to the organisation. Although being involved with the project only indirectly, their inside/outside view highlighted some elements around diversity and inclusion which were not mentioned by others. Their insights are very relevant for the integration and sustainability of the project within the organisation. They were used to attending external events with similar organisations and exchanging ideas with former colleagues. Those spaces were inspiring, and the participant could make connections and draw some interesting ideas.

Noticeable in terms of power and agency was the balance and dynamics between students, organisation, and participants. Many participants mentioned a growing awareness of the value of the data and an increased appetite for risk within the organisation. Respondents were then presented with the hypothetical possibility of extending the ML capabilities to include a performance management tool to be used on staff. This produced different reactions. Permanent staff were surprised. Some could not see the benefits of doing something similar to the project being currently developed. When presented with similar reasons to the ones used for the LA project (monitoring the work and predicting performance in order to better support staff while they were doing their work) permanent staff usually responded quickly, saying a similar system applied to staff would be strongly opposed by internal resistance and unions, that would resist both monitoring and prediction. Temporary staff/consultants were more open and could see the real possibility of using ML on staff.

Yet, when asked if they would accept having their own work or location tracked, a stronger response was observed, paired with some reservations and less inclination to accept. The role-reversal scenario highlighted the power imbalance between permanent staff, consultants and students, and different degrees of protection.

2. Case Study 2 (CS2)

2.1 Documents

In the following section, some key documents, including the DPIA, promotional material, and privacy policy documentation will be presented. Participants considered these documents important for conveying a comprehensive view of their approach.

a. DPIA

DIP performed 2 DPIAs. The first was mainly focused on risks and data retention. The company was mitigating the level of risk by reducing the individual's personal data retention time (maximum of 30 days). They planned to create another DPIA while deploying the portal for their customers (organisations) and the data trust. They were envisaging an increased level of risk resulting from more complex relationships between stakeholders: company, organisations and individuals; company (commercial)-company (data trust) (individuals' data are formally processed by two different legal entities); company (data trust) and users (user data can be deposited by both user and other organisations.

The relationship between DIP/data trust and third parties was not mentioned. The most recent DPIA is a long and detailed document about the data trust service. The document was created by the company following the ICO's indications on DPIAs (ICO, 2020a). The need to create a new DPIA had different reasons:

- The use of AI/ML for security monitoring purposes and real-time alerting for potentially fraudulent activity.

- The combination, comparison or matching of data from multiple sources to validate the identity.

- The use of federated identity assurance services (which link identities across different identity systems). Of interest is the clarification given with regards to the need to perform a DPIA for the Federated identity assurance. The ICO was reported as having identified this practice as always requiring a mandatory DPIA (which raises a question about the lack of a similar recommendation in the case of AI/ML).

DIP seemed to be willing to make the requirement to perform a DPIA more stringent, as their policy dictates that major systems processing personal data will always be subject to a DPIA, either before the implementation or following a significant change.

The ICO does not seem to have been consulted with regard to the DPIA.

Legal bases for processing personal data. DIP/data trust has a legitimate interest in processing data to provide services to organisations. The consent of the individual account holder is the legal basis for processing when data is passed on during login, and every time the document is requested (if this is required according to its level of security settings). Consent can be withdrawn, and organisations are informed by DIP with routine reporting.

Purposes of Processing. Validating the identity of the individual; accessing data trust services; accessing organisation services; checking if the documents have expired; responding to queries; improving services; protecting from malicious and fraudulent uses.

Process and data. Individuals can register with the data trust using an unverified account, or using a verified identity via a federated ID or the App. An anonymous token is generated. Data obtained for a verified account is sent to third parties for checking against their records.

Different kinds of checks can be performed, according to the identity standards defined by the government and industry, and only after receiving consent from individuals.

Data. The data trust stores the data the individual produces in their interaction with the organisations. Data is gathered via various sources: individuals, organisations, third parties, and from the device used to access the service.

In general, the types of personal data to be checked for identity validation are:

ID documents with pictures; picture/video; documents from authorities (such as Council Tax); checking with their Local Authorities if individuals have recently used their services; social media; bank/financial details. Organisations can require the individual to provide further data according to their specific needs.

The individual "*chooses who has access to their data and when*" (DPIA). Checks are performed only after receiving consent from individuals. If this is not given, a verified account may not be offered. Organisations may ask individuals to provide additional information or deny the service.

Records of these interactions are stored on blockchain.

Different levels of validations and different ways to access the services are offered by the data trust (e.g., via the App, federated ID or via a human who validates the identity).

Data is usually deleted after the checks, and in some circumstances, DIP has to retain some information for audit purposes. Data stored in the data trust is regularly checked (e.g., expiry date).

The setting-up of multiple personas gives the individuals the possibility to interact with more than one organisation without disclosing their real identities. All personas are "accountable and traceable" to the individual's identity via the data trust.

Other personal data, as per GDPR, is collected from the use of the system: the device type; the unique device identifier (e.g., the IMEI number of a mobile phone); the operating system and browser versions; and the IP address, used by DIP to identify the location of the individual.

According to the DPIA, no special categories of data are processed.

Transparency and data subject rights. Individuals can see the services with which they have enrolled, the optional information they have agreed to share, and can manage and revoke their consent from the portal. Data can be accessed, updated, amended, or erased. Data is deleted after 30 days. However, to delete the information held by organisations individuals must directly contact the organisations to request further action. DIP "can provide a list of organisations with which they have enrolled" (DPIA).

Security. Data is encrypted and no data is transferred outside the EU.

Management: The Data Sharing Agreements are approved and managed through a formal change management process by the DIP/data trust which meets periodically. "The project board assesses all changes for risk, impact, supportability, privacy and alignment with architecture principles, standards, relevant legislation and Government policy guidance" (DPIA).

Accountability and stakeholders. Stakeholders have a different degree of accountability at various stages of the process.

1. When the individual requests an account via organisations (such as the federate IDs),

   organisations are the controller of the data provided by the user
2. When the account is created and the data is transferred into the data trust,

   DPI/data trust becomes the controller.
3. When organisations request access to data held in the data trust, and individuals give consent, organisations and DIP/Data trust become joint controllers.
4. When third parties are used by DIP to validate the identities,

   they are processors for the data trust (controller) of individuals' data.

Noticeable is the only part of the document directed at the relation between DIP and its third parties. A few lines include the names of the 4 companies performing different types of searches, their degree of accountability (data processors), and their "explicit GDPR compliant schedules" (DPIA).

b. Promotional brochure (DIP commercial services and data trust)

The analysis of the promotional brochure provides an interesting insight into how the organisation presents the project externally. The project is presented as personal data exchange, a solution able to solve organisations and individuals' different needs, reducing compliance costs, increasing privacy, transparency, control on data, and inspiring social change. The Data Trust is also presented as an answer to the government request to improve access to open data to encourage the growth of the AI industry in the UK.

Compliance, fairness, transparency, and protection of individuals' rights are clearly stated in the company's mission, next to a pledge to support the government and the UN policies around Sustainability. The company was looking at launching the project in 2020 and they were looking for partners sharing the same values. Amongst other qualities, they offer transparent governance based on ethical principles. The Data Trust created by DIP provides services to individuals and pays a return to members. DIP Commercial Services generate revenue based on "an individual's ongoing consent". The personal data processed by DIP can be various, such as, biometrics, medical and financial data, immigration status, financial data. Revenues are generated by subscription, transaction fees and advertising. The document clearly states data is not provided to third parties for sales, marketing, or research purposes. Commitment to compliance and open data is provided by the willingness to engage regularly with both the ICO and the Open Data Institute. The Data Trust is managed via collective governance of members and organisations. Individuals can "invest" their data in an anonymised and aggregated data portfolio and gain a return based on the amount of data shared and the profit generated by the portfolio.

c. Privacy policies

Two privacy policy documents were shared with the researcher:

1. The applicant privacy policy: to be signed by the applicant before providing data while dealing with organisations (e.g., onboarding process). Applicants are reassured they are in control, and they provide their consent by taking a selfie and clicking a box. This confirms the consent for the data and photograph to be used to conduct the checks requested by the onboarding organisation. Collected data can vary and they may be asked to provide or confirm any social media profile details. Data is shared only with third party data and Service Providers for the purpose of completing the service checks requested by the organisation. It is not provided to third parties for sales and marketing or third-party research.

After verifying the identity via an automated process, a certificate with the results is sent to both the organisation and the applicant. If there are any inaccuracies in the data, the applicant should let the organisation know immediately. The applicant is also provided with the contact details of each of the third parties who have provided the results. This gives applicants the possibility to contact them directly concerning the results provided. The certificate and data are kept for 30 days. If criminal intent or inappropriate conduct is suspected, data can be retained for longer. Biometric information can be retained for a period of seven years as part of the crime prevention service.

The right to restrict further processing of the data can be requested by the applicant to the organisation. Applicants can object to the performing of the due diligence checks and will need to agree on an alternative course of action with organisations.

Third-party may retain a record of the enquiry for technical monitoring, service quality improvements, troubleshooting and billing purposes, but cannot use data for any other purpose.

SARs are satisfied by sending the certificate and no other info is provided to data subjects.

2. The user privacy policy: to be signed by an employee of the organisation using the App.

Data is requested for specific reasons (e.g., security, prevention of abuse, customer service and research), is kept for a year after the end of the employment and is protected by technical and organisational measures. SARs can be satisfied by accessing the portal or via the website.

Both documents provided a unique insight into their understanding of DP and AI.

2.2 Observations

This section contains some reflections on the interaction with the participants.

The cooperation with DIP was particularly positive. The CEO was the researcher's main contact. DIP demonstrated since the beginning the desire to help with the research. The nature of the research, and its focus on the GDPR, AI and FAT Principles were all of interest to the company, as these were also important elements in their project. Their participation was seen as an occasion to gain a better understanding of how to improve the project and their services in terms of data protection.

The participants were engaging with a series of organisations and institutions, and these contacts reinforced the idea that their project was fairly unique within the digital identity market. They kept me informed about some meetings they had with the Open Data Institute and their plans with regard to the ICO. The communication with both participants was easy. The CEO was always very attentive in interacting via email, finding convenient moments for meetings, providing documents I had requested and others I was not aware of. Both participants had read some texts on AI and biases, were interested in social change and were inspired by the Cooperative culture and the Quaker social activism.

The company had recently moved into a new space, a building in the city centre of a big UK city recently converted into a supporting hub for start-ups and co-working space. Both interviews were conducted in the café inside the building.

During the first introductory chat with both DIP members, we realised we all had worked for the same financial institution. The common element in our background created the possibility to make connections between the work ethics and practices in that environment and their work. Our shared experience was referred to during the interviews to better explain some elements of their current project. The CEO was willing to talk to me about what they had done, and their current and future plans. He was willing to share some documents and facilitate the interview with the second participant.

Their desire to use the project to do good and to facilitate social change was a constant element during the interview. They had incurred a few issues in creating the start-up and in promoting their project. They were also aware of the need to be part of a different environment, and of the impact the uncertainties caused by Brexit could have on the project.

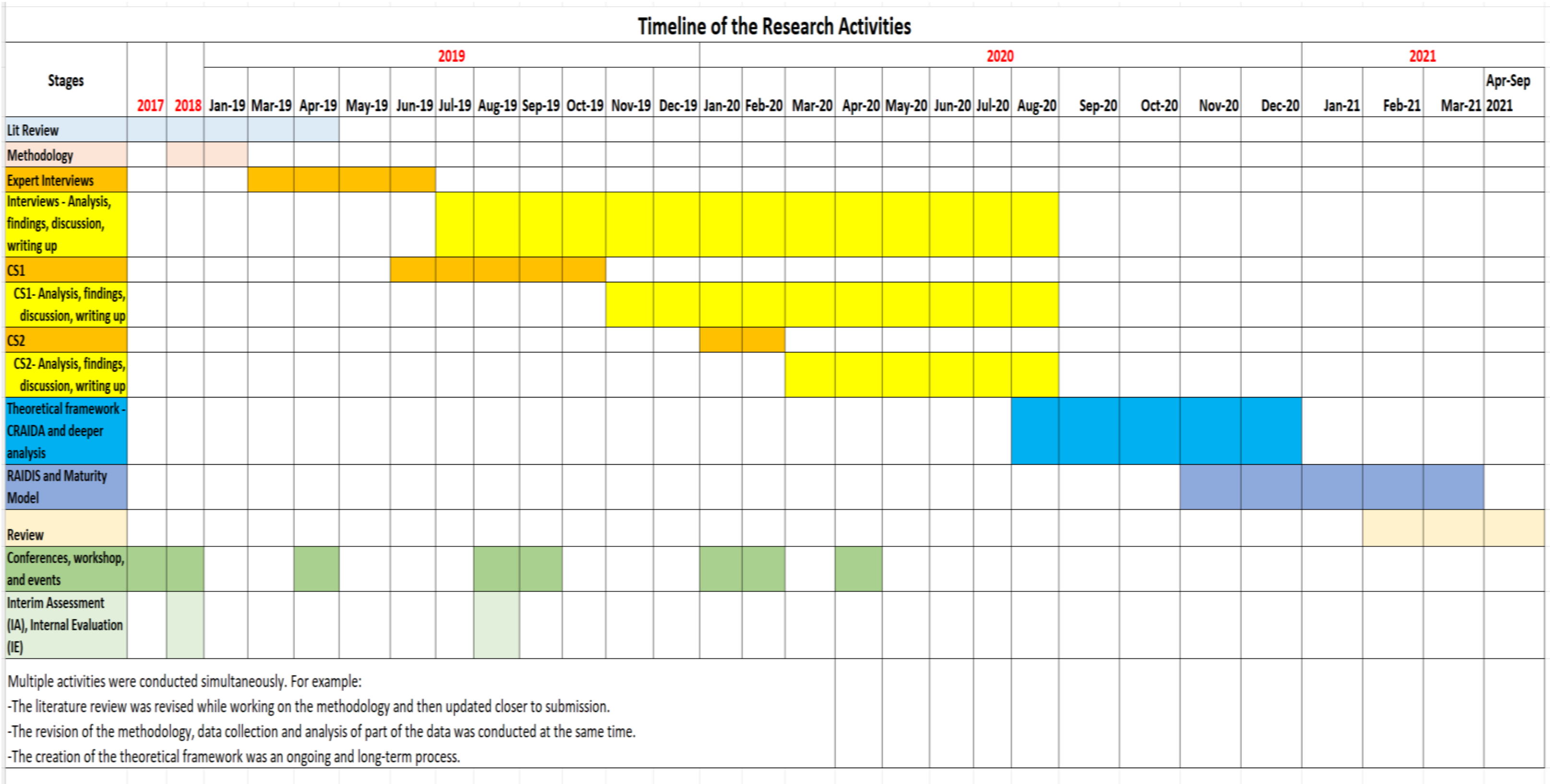
However, creating a new business model able to facilitate the empowerment of individuals to control their own data, while also facilitating social change, appeared to be a strong motivational factor for the participant. This emerged during the interview and in the communication prior and post-interview.

The second participant knew less about the research and did not know what to expect from the interview. The conversation was again very informative and pleasant. They reported having read extensively on AI and biases and showed awareness of intersectional dynamics. They believed empowering people was necessary to respond to the growing power of organisations exploiting personal data. The participant was happy to share their opinion on various aspects related to AI and data protection, and the interview was rich in details of business practices and ethics.

Both participants were available for further questions after the interview.

Appendix C: Timeline of the research activities

Table C-1 Timeline of the research activities



Timeline of the Research Activities

Stages			2019											2020												2021			
	2017	2018	Jan-19	Mar-19	Apr-19	May-19	Jun-19	Jul-19	Aug-19	Sep-19	Oct-19	Nov-19	Dec-19	Jan-20	Feb-20	Mar-20	Apr-20	May-20	Jun-20	Jul-20	Aug-20	Sep-20	Oct-20	Nov-20	Dec-20	Jan-21	Feb-21	Mar-21	Apr-Sep 2021
Lit Review	■	■	■	■	■																								
Methodology		■	■																										
Expert Interviews				■	■	■	■																						
Interviews - Analysis, findings, discussion, writing up								■	■	■	■	■	■	■	■	■	■	■	■	■	■								
CS1							■	■	■	■	■																		
CS1- Analysis, findings, discussion, writing up												■	■	■	■	■	■	■	■	■	■								
CS2														■	■														
CS2- Analysis, findings, discussion, writing up																■	■	■	■	■	■								
Theoretical framework - CRAIDA and deeper analysis																					■	■	■	■	■				
RAIDIS and Maturity Model																								■	■	■	■	■	
Review																											■	■	■
Conferences, workshop, and events	■	■			■				■	■				■	■		■												
Interim Assessment (IA), Internal Evaluation (IE)		■							■																				

Multiple activities were conducted simultaneously. For example:
-The literature review was revised while working on the methodology and then updated closer to submission.
-The revision of the methodology, data collection and analysis of part of the data was conducted at the same time.
-The creation of the theoretical framework was an ongoing and long-term process.