Image by Ana Rojas Silveyra

The Challenges of Big Data Analytics on Responsible Decision-Making in Asymmetrical Relationships: A Forward-looking Approach to Responsibility

Juan Diego Muñoz Arbeláez

Master Thesis

To obtain the degree of

MSc Philosophy of Science, Technology, and Society

Faculty of Behavioral, Management, and Social Sciences

Supervisor: Dr. Dina Babushkina

Second Reader: Dr. Julia Hermann

University of Twente

Enschede, the Netherlands

August 2023

Acknowledgments

The ones that have had to put up with my inconsistent mood and insecurities know this process was challenging for me. This only happened because of the support, confidence, and love of the people around me.

I would like to thank Dr. Dina Babushkina for her unconditional supervision, guidance, confidence, and feedback.

I would also like to thank Dr. Julia Hermann for her great predisposition to help at different stages of this thesis.

Thanks to my mother for always being strong, independent, and supportive.

Thanks to Roos, Victor, and Luca. You made my life especially joyful during the last year and were always willing to talk about anything. I will miss you.

Thanks to Sam. I have learned many things from you, but especially how to be strong and move forward. Thanks for the talks about life in general and this thesis in particular. I love the house I ended up living in.

Thanks to Marco for his support during the months we had the pleasure and joy to share.

Thanks to the cube people: Dami, Alexis, Nina, Pietro, and Gemma. I am looking forward to a year of adventures with you.

I learned how not to write a thesis. I have faced my anxiety when it comes to crafting a decent text. Now, I think I have a better idea of what this is about, only if I could write it again.

Table of Contents

Summary

The main research question of this thesis aims to investigate the challenges of Big Data Analytics for responsible decision-making in asymmetrical relationships. It is divided into three sections, each addressing specific sub-questions.

The first section examines some of the challenges Big Data Analytics poses, focusing on the obstruction to providing relevant explanations within asymmetrical relationships in decision-making processes, and it finishes by posing the question of how the ability to provide explanations is a problem for responsibility in the philosophical debate surrounding the ethics of AI. The oversimplification of complex social issues through data-driven algorithms reduces the ability to understand and provide meaningful explanations. Addressing social issues mainly with technical solutions will have major limitations and shortcomings. Additionally, concerns about inscrutable and inconclusive evidence point to problems with attribution of responsibility and the ability to provide explanations. However, beyond attribution of responsibility, this impacts the moral patient's empowerment and the moral agent's sense of responsibility, diminishing trust between them. The implications of poor explanations in asymmetrical relationships have consequences for social cohesion and responsibility beyond questions of attribution.

The second section delves into moral responsibility, adopting a forward-looking perspective that incorporates relational and interpretative elements. This section aims to understand what responsibility means and what notions of responsibility are relevant for decision-making. It begins by looking at how responsibility is discussed in the field of ethics of AI to address the concerns posed by Big Data and AI and then moves to propose a forward-looking approach. This approach adopts a feminist perspective and recognizes the interconnectedness and interdependence of individuals in decision-making processes. It emphasizes the importance of explanations for care, trust, and understanding in asymmetrical relationships. Moreover, it seeks to involve the moral agent and the moral patient during the decision-making process.

The third section explores the analogy of industrial maintenance practices to show how caring by explaining could look in practice. The aim of this section is to show how responsibility can be enacted through explanations in practical settings. Maintenance practices emphasize the need for meaningful explanations to sustain relationships based on reliability and trust.

The research methods used in the thesis include a literature review, conceptual analysis, and analysis of practical cases. The literature review provides a basis for understanding some of the challenges and implications of Big Data Analytics in decision-making processes. The study of practical cases examines real-world scenarios where the impact of Big Data Analytics on asymmetrical relationships is evident, emphasizing the importance of relevant explanations. The conceptual analysis offers insights into moral responsibility from a feminist perspective, highlighting its relational and interpretative dimensions.

This thesis shows that Big Data Analytics complicates the ability to provide meaningful explanations in decision-making, challenging different notions of responsibility. The oversimplification of complex situations through data-driven algorithms hinders the capacity to understand and provide reasons for decisions, exacerbating power imbalances in asymmetrical relationships. Exacerbating these power imbalances is problematic for responsibility beyond attribution because it reduces reliability and trust in human relationships necessary for social cohesion and stability.

The analysis argues for a forward-looking approach to responsibility that encompasses care, trust, and understanding. Actively involving moral patients in decision-making and recognizing their requirements for meaningful explanations can foster trustworthy and reliable relationships.

The major conclusions drawn from the research highlight the need for a critical engagement with the challenges of Big Data Analytics on responsible decision-making. Decision-makers must acknowledge the limitations of algorithmic outputs and prioritize providing explanations that go beyond reasons based on accuracy and efficiency. A forward-looking understanding of responsibility based on the notions of virtue, moral obligation, and

answerability helps stress the need for relevant and meaningful explanations during the decision-making process and not only after the fact. Relevant and meaningful explanations are important because, through them, trustworthy and reliable asymmetrical relationships can be achieved. Therefore, the main aim of caring by explaining is to foster trust and reliability in asymmetrical relationships. Acting responsibly means looking forward to maintaining high degrees of trust and reliance within these relationships.

This thesis highlights the importance of relevant and meaningful explanations to maintain trust and reliance in asymmetrical relationships. By embracing a forward-looking perspective of responsibility, inspired by feminist theory, and incorporating insights from industrial maintenance practices, it proposes a forward-looking approach to responsibility that prioritizes care and understanding. The research encourages decision-makers to actively involve moral patients in the decision-making process, promoting responsible decisions that protect the vulnerable and improve social cohesion.

Introduction

Human decision-making is a complex and intricate process. Justifying and explaining our decisions can be problematic, especially in a society that increasingly relies on data and technology to support decisions in many aspects of human activity. The increasing use of Big Data Analytics to inform decisions has brought new challenges for understanding and maintaining responsible decision-making in asymmetrical relationships. Decision-making in several domains, such as courts and healthcare, has been affected significantly by technologically informed decisions (Green & Chen, 2019; Zhou et al., 2019).

Decision-making within complex social and technical systems[1], such as healthcare or governance, is influenced by a framing, narrative, and socio-political context. In decision-making processes that have societal relevance, there is often someone in a vulnerable position; a person affected by the decisions made by someone else. In ethics terminology, these can be called the moral patient (vulnerable position) and the moral agent (decision-maker). The relationship between the moral agent and the moral patient forms a unity. These relationships can be asymmetrical, characterized by an imbalance of power where the moral agent has the authority to decide upon a course of action that will significantly impact the moral patient—for instance, the relationship between lawyer/client, patient/doctor, or police officer/citizen, where the expertise and knowledge of the decision-maker (moral agent) provides guidance or decides and has authority over the moral patient. However, these asymmetries are not necessarily problematic; they manifest in different degrees and can be inherent to a specific role. Asymmetrical relationships are, therefore, crucial for social cohesion because they contribute to the system's functioning and stability.

The dependency of the moral patient raises questions about the extent of legal, professional, and moral responsibilities that the moral agent should have. Professional responsibilities may serve as a guideline for individuals in positions of power. However, legal and

[1] System composed of social (individuals and social institutions) and technical elements to achieve a function in society (Geels, 2004; Nickel et al., 2010).

professional responsibilities, such as physicians' ethical codes or consent forms, may not be sufficient to establish and maintain trusting and reliable relationships between the moral agent and the moral patient. Therefore, moral responsibilities become relevant in asymmetrical relationships where vulnerability is key.

Moral responsibility goes beyond legal and professional obligations. It encompasses a broader sense of empathy and ethical engagement. Moral responsibility acknowledges the moral agent's duty to act in the best interests of the moral patient, considering their perspective and needs. It requires the moral agent to go beyond compliance and reliance and actively engage with the moral patient through the decision-making process.

Big Data Analytics refers to the application of Artificial Intelligence (AI) algorithms to analyze large amounts of data to inform or aid decision-making. Big Data and AI algorithms depend on each other materially and conceptually. Materially because data is created and processed to feed AI algorithms. Conceptually, I aim not to cover these terms extensively but to present an overview of their underlying assumptions and show how they layer up and work together. Moreover, I am interested in how AI algorithms affect asymmetrical relationships and decision-making in practice. This means I will focus on the decision-maker (moral agent) and the person affected by the decision (moral patient) rather than on the designers and developers. In practical settings, I want to explore how the way we talk, think, and develop an understanding of technological applications impacts their use for decision support and, in turn, affects human relations. While the technical aspects of Big Data and AI algorithms are relevant, they are not my main concern.

Within AI, I will focus on Machine Learning algorithms because their alleged capacity to adapt and extrapolate patterns makes them relevant for extracting valuable information from large amounts of data (Big Data). Furthermore, I will use the term algorithms to refer to computational techniques or mathematical constructs implemented and configured through technology for a specific task following Mittelstadts' et al. (2016) definition. In this sense, I am particularly interested in the application of algorithms that integrate into complex social and technical systems informing decision-making processes. However, I will only discuss applications where a human agent makes a final decision regardless of the

possible analysis or recommendation of algorithms. This means that concerns about automated decision-making, autonomy, and moral agency are beyond the scope of this work.

The main question of this thesis is: what are the challenges Big Data Analytics poses for responsible decision-making in asymmetrical relationships? To answer this question, this thesis is divided into three sections, each one addressing specific sub-questions.

The first section aims to set the stage for the discussion. It explores how Big Data Analytics influences decision-making within asymmetrical relationships and challenges notions of responsibility. It is driven by two sub-questions: What is its context of deployment and implementation of Big Data Analytics within asymmetrical relationships? Why are the challenges Big Data Analytics poses for providing explanation a problem for responsibility? The main objective is to understand better how the inability to provide an explanation is a challenge for responsible decision-making in asymmetrical relationships.

Big Data Analytics hinders the ability to provide relevant explanations within asymmetrical relationships, reducing trust and impeding the moral patient's empowerment. As I examine the implications of hindering the ability to provide explanations for responsibility, it becomes clear the focus on the attribution of responsibility, often from a backward-looking perspective. However, explanations are valuable not only to determine who is responsible but are critical in building trust and empowering the moral patient as they offer a means for self-protection and reduce power imbalances. Big Data Analytics should not overshadow the importance of relevant explanations. This leads to question how the lack of explanations is understood as a problem for responsibility and why.

The second section starts by exploring the current debate about responsibility in the field of ethics of AI. This section aims to answer how are the main challenges posed by Big Data Analytics regarding the capacity for providing explanations addressed in the current debate and how we can think about them differently- taking a forward-looking approach to responsibility.

To explore the concept of responsibility in the context of Big Data Analytics, I start with the current debate about “responsibility gaps” in the field of ethics of AI (Matthias, 2004). The

debate about responsibility around AI often takes a backward-looking approach focused on the attribution of responsibility. With this in mind, this thesis aims to explore a forward-looking approach that puts aside questions about attribution of responsibility. In doing so, it adopts a feminist perspective that recognizes that responsibility cannot be understood in isolation but within the network of relationships and socio-political context. It considers cooperation, connection, individual experience, and vulnerability as crucial aspects of responsibility. Feminist theory highlights the interdependence between individuals and the impact of our decisions on others. It seeks an understanding of responsibility beyond individualistic notions, considering how we both influence and are influenced by others. Within this framework, responsibility extends beyond assigning blame and instead focuses on promoting care, empathy, and accountability (Adam & Groves, 2011; Shafer-Landau, 2018; Walker, 2006).

From within this feminist perspective, I propose a forward-looking approach that builds on the notions of responsibility as virtue, moral obligation, and answerability (Coeckelbergh, 2020; Richardson, 1999; Van De Poel, 2011; van de Poel et al., 2015), connecting it to reliance and trust using Walker's (2006) concepts of "trusting relationships" and "default trust." Mainly, I engage in a conceptual analysis of responsibility from a feminist perspective, trying to understand better how it relates to reliability, trust, and care. By recognizing the importance of explanations, interpretation, and relationships, the goal is to preserve reliability and trust in asymmetrical relationships, the value of explanations for the moral patient, and increase the sense of responsibility for the moral agent. However, from a more practical side, the question of what it means to act responsibly is still open.

Finally, the third section draws on my personal experience with industrial maintenance practices to show how caring can be practiced by explaining. The goal of this section is to answer the question of what it means to act responsibly in decision-making- taking a forward-looking approach to responsibility. Responsibility, viewed from a forward-looking perspective as relational and interpretative, involves the ability to provide and understand meaningful explanations. This section aims to show why explanations are relevant and how caring by explaining could look in practice. The connection between responsibility and

maintenance comes from the practice of caring by explaining decisions in relevant terms, which helps maintain reliable and trusting relationships. By answering this question, we can draw some insights from industrial maintenance to show that by providing meaningful explanations, reliability and trust can be fostered in asymmetrical relationships, protecting the moral patient and enhancing the sense of responsibility of the moral agent. Approaching responsibility from a forward-looking perspective might help us reframe and mitigate the challenge of a lack of explanation that Big Data Analytics poses.

Section 1: How do Big Data and AI influence decision-making?

In this section, I explore contextual factors of Big Data Analytics that might be relevant for decision-making. I introduce Big Data and Artificial Intelligence (AI) with in a specific narrative. Then, I use two examples to describe this context, show what are asymmetrical relationships and how they are affected by introducing Big Data Analytics. The first example discusses AI-Computer Aided Diagnosis (AI-CAD), and the second Predictive Policing. Next, I move to describe data-related and AI-related problems for decision-making and how they relate to responsibility. I argue that algorithmic analysis of data is embedded in a reductionist narrative that is problematic when applied to trying to solve social and political issues. In this context, algorithmic support affects decision-making by hindering the capacity to provide relevant explanations in asymmetrical relationships.

Let us now explain the connection between Big Data and AI in order to show how they depend on each other materially and conceptually.

Big Data and Artificial Intelligence: Big Data Analytics

Digitalization, datafication, and artificial intelligence (AI) are changing how we relate to others, the world, and ourselves. Digitalization and datafication refer to integrating digital technologies in society that allow for an understanding of the physical world, processes, and practices in terms of digital data (Southerton, 2020). For example, public transportation or banking services that previously used physical maps and cash have moved to fully digital systems we interact with through a phone or a computer. This move to digital systems has

generated excessively large amounts of data framing the world in terms of quantifiable data waiting to be exploited for economic, social, or political purposes.

In our highly digital society, large amounts of data are continuously generated, collected, and analyzed to draw patterns, correlations, and insights across several domains. Society generates massive amounts of data, also known as *Big Data*. Big Data refers to enormous datasets that are collected rapidly and combine a diverse and broad range of variables from various sources (Kitchin, 2014). The data sources are infinite, from online purchasing history, search preferences, and health information to screen time and geo-location information. Public and private organizations collect and analyze Big Data to generate insights. The analysis of Big Data to produce insights or evidence is referred to as *Big Data Analytics or Analytics*. "Mining and extracting meaningful patterns from massive input data for decision-making, prediction, and other inferencing is at the core of Big Data Analytics" (Najafabadi et al., 2015, p. 2). Big Data analytics uses computational techniques to draw patterns from data to extract value for different purposes (Grindrod, 2014). This has become one of the main objectives of businesses and governments, where Artificial Intelligence techniques, such as Machine Learning, have appeared as a perfect solution.

Artificial Intelligence (AI) development is a movement driven by the aim to understand intelligence and also build "intelligent" systems (Russel & Norvig, 2021, p. 1). In Computer Science, the term Artificial Intelligence is broadly used to refer to different technologies and methods and has multiple definitions. While there are other definitions from other disciplines, I argue that approaches stemming from computer science have contributed to a problematic narrative surrounding AI, which I will describe after unpacking some of the approaches from computer science.

Wang (2008) presents five approaches in the context of AI research as a branch of Computer Science: structure AI, behavior AI, capability AI, function AI, and principal AI. Structural and behavioral AI compare machines to human beings in terms of brain structure or mind operation. Capability and function AI focuses on practical problem-solving abilities. While capability AI considers problem-solving more generally, function AI looks for a function that

relates inputs to outputs as a part of the process needed to solve a problem. Lastly, principle AI looks at finding the best solutions given certain conditions.

Similarly, Russel & Norvig (2021) describes four approaches to AI: The Turing[2] Test Approach (acting humanly), The Cognitive Modeling Approach (thinking humanly), The "laws of thought" approach (thinking rationally), and The Rational Agent Approach (acting rationally). To "act humanly," according to Russel & Norvig (2021), a computer should possess: Natural Language Processing, Knowledge Representation, Automated Reasoning, and Machine Learning. Machine Learning is defined as the capacity "to adapt to new circumstances and to detect and extrapolate patterns" (Russell & Norvig, 2021, p. 2). This seems to be a reason why it is appealing for Big Data applications. Nevertheless, "acting humanly" is also related to "acting rationally." In the sense that humans need reasoning, knowledge representation, language, and learning to function in complex societies. This means that Machine Learning is part of Artificial Intelligence attached to at least two different approaches.

Like Artificial Intelligence, Machine Learning is a term that has multiple definitions. Bell (2022) defines Machine Learning as a branch of Artificial Intelligence that designs computational systems that can learn and improve with experience generating a model that can be used to predict outcomes. Mittelstad et al. (2016) quote Otterlo (2013), saying that Machine learning is ''any methodology and set of techniques that can employ data to come up with novel patterns and knowledge and generate models that can be used for effective predictions about the data'' (Otterlo, 2013). For Jenga et al. (2023), Machine Learning is a branch of AI that includes methodologies and techniques that use data to produce patterns that are used to predict future outcomes or behaviors. Deep learning, also called Deep Neural Networks, is an example of a Machine Learning technique. Within Machine Learning, Deep Learning (Deep Neural Networks) are algorithms "largely motivated by the field of artificial intelligence, which has the general goal of emulating the human brain's ability to observe,

[2] For more on the Turing Test, see Turing (1950)

analyze, learn, and make decisions, especially for extremely complex problems" (Najafabadi et al., 2015, p. 4).

Machine Learning algorithms can be categorized into two learning types: supervised and unsupervised (Bell, 2022). Supervised learning refers to the use of labeled training data. This means there is a known correct output for every training data input fed into the algorithm. On the other hand, unsupervised learning uses input data without a known output. With unsupervised learning, "you let the algorithm find a hidden pattern in a load of data. With unsupervised learning, there is no right or wrong answer; it's just a case of running the Machine Learning algorithm and seeing what patterns and outcomes occur." (Bell, 2022, p. 211).

In this context, their creators envisioned Machine Learning algorithms as a perfect solution to draw insights and meaningful correlations from large and complex datasets. However, some of these techniques challenge human capabilities for different reasons, for example, because, even if reliable, their complexity, internal processes, and logic are incomprehensible to us to a certain extent. Examples of problematic use of AI in decision-making can be encountered in different domains, for example, Human Resources management and criminal justice. In both cases, a moral agent in a position of power (judge/human resources manager) makes a decision that affects the moral patient (offender/job candidate) with the support of AI algorithms. Moreover, other challenges stem from how we describe Big Data and Machine Learning algorithms. A lack of clarity about the limitations and expectations of AI in these approaches contributes to a narrative about AI that can be misleading and exacerbate challenges posed by this technology. Let us discuss what this narrative is and how it is problematic.

A Narrative about AI

When using Machine Learning algorithms, the decision maker is embedded in a way of thinking about the world, intelligence, and artificiality. Framing the world in terms of data and using AI algorithms to analyze it, reduces the human cognitive processes associated with intelligence and gives the output of algorithms unjustified authority over human analysis. It

seems we are trying to remove the human factor to overcome the problems related to data processing while retaining the capacity to derive, assign, and justify meaning in patterns. By doing this, meaning that was allegedly in data is extracted by algorithms and given to us.

Defining intelligence is complex; every attempt to define it delimits and determines what is and what is not. In general terms, computer science claims to develop "intelligent" systems where the meaning of intelligence takes different shapes. A vague narrative of AI capable of performing tasks that require intelligence without the limitations of human beings seems common in public discourse. These narratives do not define what they mean by intelligence, implicitly reducing the concept to one of its aspects or including in the definition other terms, such as thinking and self-awareness. For example, TechRadar.com says, "AI as a concept refers to computing hardware being able to essentially think for itself, and make decisions based on the data it is being fed" (Moore, 2019). There is a risk in moving too fast from intelligence to learning to understanding of a machine. For example, an article about AI on MIC.com says: "But the most important core component of AI constructs and programs, given that they're modeled after human intelligence, is the fact that they learn. In fact, they'll display some of the same behaviors as humans when they're beginning to understand something" (Vincent, 2019).

This narrative is problematic because it creates an imaginary of something artificial but intelligent in the same sense as humans. Portrayed as being modeled after human cognitive processes or structures such as learning or brain neural networks, these data-driven algorithms have gained validity and popularity in domains such as health diagnostics, human resources management, and governance of public affairs (Green & Chen, 2019; Miotto et al., 2016). The necessity to replace humans with artificial intelligence arises from their inability to process vast and intricate datasets. In this case, intelligence is understood mainly as data processing and pattern recognition. This reduction of the meaning of human intelligence wrongly implies that removing the human factor entails no major problems. In other words, Machine Learning algorithms seem to outperform human beings in data processing and interpretation without any problem. However, there is more to human cognitive processes related to intelligence than (machine) learning. The framing of Big Data Analytics combined

with a narrative where AI can outperform humans' data processing and cognitive abilities gives an unjustified autonomy and authority to the outcome of those algorithms.

For instance, concerns about misrepresenting data are exacerbated by a utopic narrative of AI. Imagine a perfect trend between car accidents and ice cream sales in Amsterdam for the past three years. This association is meaningless, and the fact that both trends align does not mean that there is correlation or causation between them. If humans construct meaning around these trends, it is open to scrutiny and questioning. On the contrary, if an association or causal relation comes from an "*intelligent''* machine that is "intelligent" to the extent that it outperforms humans in a certain isolated function, it may appear believable. This is a problem because it provides more credibility to arguments that rest on unsound grounds. In other words, AI algorithms are built on "the widespread belief that large data sets offer a higher form of intelligence and knowledge that can generate insights that were previously impossible, with the aura of truth, objectivity, and accuracy" (boyd & Crawford, 2012, p. 663). This adds up to misleading discourse about "intelligent" AI algorithms that are said to be able "learn" or even "think" independently.

Additionally, the way we describe AI algorithms can contribute to normalizing and constructing meaning by defining terms related to AI and human intelligence. Terms related to human intelligence that do not have a simple, comprehensive definition have come together to describe Artificial Intelligence algorithms. For example, knowledge representation, reasoning, rationality, and learning. These terms are defined in a way that intelligence is associated mostly with information processing, prioritizing and normalizing one way of understanding intelligence. This normalization entails that other possible ways of understanding intelligence are not considered creating a reduced standard of intelligence; machine intelligence. However, we should be careful about definitions and engage critically with them, being aware that they could be otherwise. Rather than arguing for the best definition, it is important to understand their differences, implications, and relations. Each definition specifies intelligence or other terms in a certain way and makes different assumptions about the system and its environment. This means that each definition frames the problem in a particular way, predefining solutions to a certain extent.

So far, I have connected Big Data and AI within a particular narrative that affects their deployment and implementation. When introduced in complex social and technical systems, Big Data Analytics affects the relationship between moral agents and moral patients. Particularly asymmetrical ones. Next, I will use two examples to further explain asymmetrical relationships, the contextual factors around them, and show how the moral patient is affected. These two examples intent to show how Big Data and AI affect relationships and the sense and notions of responsibility. I will attempt to demonstrate how AI algorithms, driven by data, present challenges for society that extend beyond technical considerations. In both, the moral agent responsible for a decision is in a position of authority, while the moral patient has some degree of vulnerability. Moreover, these algorithms are integrated into complex and socially relevant systems where stakeholders interact.

Asymmetrical Relationships, Big Data, and AI

Asymmetrical relationships are characterized by an imbalance of power between moral agent and moral patient. The imbalance can be significant due to a combination of factors such as authority, knowledge, expertise, or resources. A higher level of education, access to information, or understanding impacts power asymmetries. In these situations, the moral agent can decide upon a course of action that will significantly impact the moral patient—for instance, the relationship between lawyer/client, patient/doctor, or police officer/citizen, where the expertise and knowledge of the decision-maker (moral agent) provide guidance or decides and has authority over the moral patient. This means a position of vulnerability of the moral patient compared to the agent. However, these asymmetries are not necessarily problematic; they manifest in different degrees and can be inherent to a specific role.

I am aware that there are mechanisms to reduce power asymmetries in relationships to reduce the degree of asymmetry. In the relationship between patient and physician, the degree of asymmetry could be less compared to predictive policing. In healthcare, there are mechanisms known for giving the patient participation in the decision-making process and

save guarding her autonomy[3]. However, AI poses challenges to these mechanisms developed to protect the moral patient (Bjerring & Busch, 2021). In contrast, for predictive policing, the relationship between law enforcement officer and victim/citizen/perpetrator could have different degrees of asymmetry- more compared to patient/physician. Nevertheless, further understanding the nuances of how power asymmetries operate in asymmetrical relationships could be relevant for society.

In roles where the moral agent is in a position of providing care, such as patient/doctor or student/teacher, these asymmetries are inherent to the relationship. This means there are different degrees of emotional investment, dependency, and care between agent and patient that are not necessarily harmful, can increase or perpetuate power imbalances. However, dealing with the dynamics in power asymmetries means acknowledging and understanding the nuances of asymmetrical relationships in various contexts, from personal relationships to professional settings that affect broader societal structures. To do this, care ethics could help us by addressing the implications of these power imbalances, seeking to challenge and transform societal norms and structures, and striving for reliable and trustworthy asymmetrical relationships.

AI-CAD Example

AI computer-aided detection/diagnosis (AI-CAD) based on image processing and interpretation is an example of how Big Data and AI framing and narrative affect the relationship between patient and doctor in clinical practice. This relationship is asymmetrical because the patient is in a vulnerable position with respect to the doctor. The patient will be affected by any decision that is made and the strength of the relationship depends on the quality of explanations provided. Decreasing the quality of possible explanations is problematic for responsibility.

With this example, I want to show two problematic aspects of AI-CAD. First, the narrow framing of diagnosis through an image-driven approach leaves physicians and patients in

[3] For more on patient-centered care, see: (Kwame & Petrucka, 2021; Maeseneer et al., 2012)

the background of technology, maintaining a lack of understanding of social limitations. Second, feature determination by AI algorithms rather than humans does not allow for meaningful explanations.

CAD and AI-CAD emerge in a particular social, economic, and political context where a narrative promotes their use and legitimization. The world's first computer-aided detection (CAD) device, approved by the U.S. Food and Drug Administration in 1998, was meant to assist radiologists in analyzing mammograms to improve the accuracy and efficiency of breast cancer detection. Its commercialization in the United States was promoted in 2002 by the reimbursement for x-ray scans such as chest computer tomography and colonoscopies. However, other countries were more cautious about their application and commercialization. Japan, for example, approved CAD only for mammography in 2018 (Fujita, 2020, p. 6). In the United States, the context promoted the use of traditional CAD, reinforcing the legitimacy of using medical images for diagnosis. The proliferation of medical images, then, pressured radiologists and their interpretation and classification capabilities.

Traditional CAD systems were meant to help radiologist process and interpret medical images by working around them with a defined purpose. It was based on "how to use the circumstances surrounding the physician's image interpretation" and on their purpose (Fujita 2020, p 12). This means that attention should be paid to how the introduction of CAD impacts social dynamics in practical settings in which they are deployed. However, even though the use of CAD has been widely accepted in the United States, there were concerns about its ineffectiveness in clinical settings (Fujita 2020).

In this image-driven diagnostic context, AI appears with hopes of helping physicians in their work and fulfilling expectations of increasing the accuracy and efficiency of diagnosis. The addition of AI to CAD does not seem to provide any understanding or insights about how to improve ineffectiveness in clinical settings or other social limitations. On the contrary, it promotes an excessive focus on image processing, neglecting the importance of other information relevant to medical diagnosis (Buhmann et al., 2020).

The main difference between CAD and AI-CAD is feature definition. To help physicians with medical imaging processing and interpretation, a CAD system would search for specific features in X-ray images to classify regions and identify potential abnormalities. In traditional CAD systems, those features are determined by the developers and designers of the software (Fujita, 2020). In contrast, with AI-CAD, feature identification is performed by Machine Learning algorithms. With the introduction of Machine learning, humans do not define the features to look for in the images. Figure 1 shows how Deep Learning takes over feature definition and classification (Fujita, 2020). This is problematic because it reduces the understanding of intelligence and outsources meaning assignment to Machine Learning algorithms.

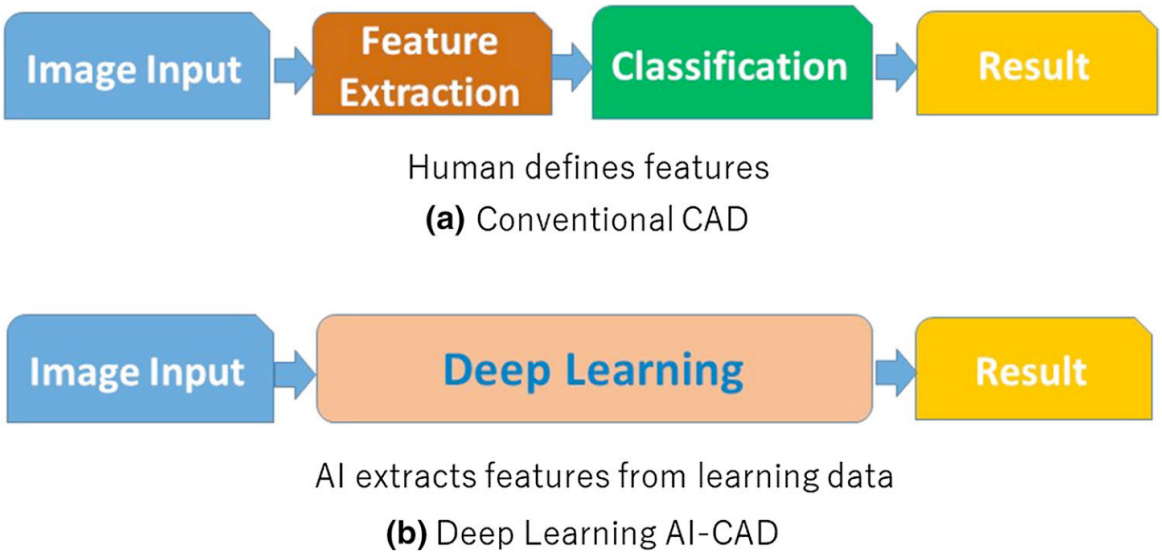

Figure 1. Traditional CAD vs. AI-CAD (Fujita, 2020, p.8)

Considering that a Machine Learning algorithm possesses some kind of intelligence similar to a person promotes a narrow understanding of human intelligence as learning and information processing. Part of the high expectations of AI stems from the understanding of it as being able to perform human cognitive functions at a large scale, rendering large amounts of data into meaningful and useful information. As mentioned above, while CAD approaches use a human designer to come up with relevant features for the images, AI-CAD creates its own features (Fujita 2020). Moreover, it assumes that an algorithm can assign

meaning to features in the same way a person does it. The translation of data into meaningful information seems to be taken for granted when algorithms are portrayed as entities that possess intelligence, can learn, or are modeled after the human brain. Reducing cognitive processes to functions, such as learning, and outsourcing meaning, challenges decision-makers' capacity to explain their decisions. Imagine the impact this has on a physician that relies on AI-CAD for a diagnosis but (1) neglects the importance of non-medical image information for diagnosis and (2) cannot fully explain why the features are relevant and how they are connected in a meaningful way.

With this, I am not saying that AI-CAD is to the detriment of medical practice or that it is not useful. Undeniably, some of these algorithms have exceptional capabilities of detecting patterns relevant to a diagnosis under certain circumstances (Fujita 2020). However, , there are relevant concerns that touch upon interpretational and relational aspects. How these devices interact within networks in clinical systems and their effects on stakeholders' relationships might determine their impact and effectiveness. A better understanding of how different stakeholders interact with the system and how this affects the relationships that hold the system together is needed. In contrast, the main goal of AI as an aid for CAD (AI-CAD) is not to address possible issues of deployment and implementation in clinical settings and among people. Instead, AI-CAD focuses on technical solutions to improve the manipulation and correlation of massive amounts of data and images.

With AI-CAD, diagnosis relies less on interpretation and social practices and more on data-driven algorithmic authority. The over-reliance on data-driven algorithmic strips off meaningful explanations even if accuracy and efficiency increase. This is problematic for physicians, radiologists, patients, and society at large because it decreases reliability and trust in relationships. Physicians may lose access to knowledge relevant to provide relevant explanations to the patient. Moreover, this could decrease the degree of involvement of the patient in the decision-making process.

Predictive Policing Example

Predictive policing is another example where Machine Learning is used for analyzing large amounts of data. In Europe and the United States, different statistical methods analyze crime data to extract knowledge in the form of patterns or trends about future crimes. Traditional police intelligence is grounded in theories such as routine activity theory and crime pattern theory (Hardyns & Rummens, 2018). These theories, along with statistical models, explain the connection between input features and outputs.

Compared to other statistical models, Neural Networks provide limited insights into the relationship between input features and output, therefore limiting their explainability. This is similar to the feature determination for AI-CAD previously mentioned, where AI algorithms draw patterns from data instead of human beings, reducing the capacity for satisfactory explanations. Pitfalls associated with attribution of responsibility arise when questions about whether the sufficiency of these predictions justifies decisions. This does not mean that algorithmic output is not relevant for decision support. However, an over-reliance on technical aid does not allow for other ways of support.

Hardyns & Rummens (2018) mention a distinction between the effectiveness of predictions made by the system and the effectiveness of the system itself (p. 213). This is relevant because to use crime analysis outcomes, the operationalization and strategy of police responses are of utmost importance. This means that for implementing predictive policing effectively, the involvement of police officers in their deployment and use is key. The system should allow police officers to use their insights responsibly, meaning at least without losing the capacity to provide a relevant explanation for the citizen, victim, or perpetrator. This takes us back to thinking about responsibility not only under conditions of control and knowledge but also as answerability.

Discussions about responsibility also question the involvement of private companies that develop the analysis software. The collection and analysis of crime data could be seen as a responsibility that the police have, which is transferred to private companies through this system. A response to this would sustain that this should be seen as an opportunity for the

police to focus on core responsibilities (Hardyns & Rummens 2018). However, there is interpretative confusion about responsibilities, and precisely who is responsible for what is unclear. Decision-makers face situations of ambiguity introduced by technology with respect to moral judgment. This shows at a societal level how the introduction of AI algorithms creates uncertainty regarding responsibility in complex systems.

Moreover, companies and media present the outcome of the analysis as a literal prediction of crime when it only provides the likelihood of an event. This shows how public discourse can shape understanding of technology and influence decision-making. Furthermore, Hardyns & Rummens (2018) provide three main recommendations for implementing predictive policing: "(1) reliable data collection, (2) clear communication between different police units and hierarchy levels, and (3) police response strategy." (p. 215). The second and third recommendations again point out the interaction between people in complex systems and how it is affected by technology.

Moreover, framing the outcome as a *prediction* has an impact on real-world operations (Birhane, 2021). This impact occurs at the level of the user. For instance, if the law enforcement officer understands the algorithm's outcome as a 'prediction of crime,' she might be more inclined to see more crime in that area. Therefore, how AI for decision support is framed in deployment and use plays a crucial role in establishing a predominant social meaning and even the measurable variables that could then be used to assess its success. If the number of arrests is higher after deploying more officers into a 'predicted' high-risk area, it might be easy to think that the AI indeed *predicted* crime. This is a dangerous assumption where interpretation, correlation, and causation might be wrongfully related. Moreover, this could reinforce a pattern where it is difficult for decision-makers to think of alternative interpretations of technology and be aware of how interpretation affects their decisions.

For example, algorithms used to *predict* areas where crimes could be more likely- to deploy police officers accordingly- are known to reinforce bias and discrimination (Birhane, 2021). This ex-post analysis reveals how algorithm-supported decisions can sustain bias from law enforcement officers, creating a feeling of unfair accusations (Tollon, 2022). Furthermore, it suggests how the societal and individual meaning of technology might affect the decision-

making of police officers. From the perspective of the individual officer, there could be multiple interpretations of what the technology is and what it is supposed to do in certain contexts. For instance, if the algorithm is understood as a more reliable and epistemically valid way to determine police presence, the suggestions might go unquestioned and considered superior to what a human being can do.

This affects the moral patient by reducing the quality of the explanations that could be already at stake in asymmetrical relationships. For law enforcement officers, providing relevant explanations might not be a priority to provide explanations to victims or perpetrators, which is a problem for responsibility. This asymmetry could be exacerbated if the officer is not able to provide an explanation because the reasoning has been outsourced to technology.

In short, stemming from Big Data assumptions and Machine Learning algorithms, data is preferred over other types of information, and unjustified validity is given to algorithmic output. Additionally, embedded in the narrative surrounding AI, authority might be wrongfully ascribed to patterns and algorithmic output. These concerns highlight the relevance of understanding the role of meaning in decision-making. Taking for granted the validity and authority of the output of algorithms is problematic because we cannot explain why those correlations are meaningful. In other words, using Big Data Analytics has an implication for decision-making because it displaces meaning attribution from humans to AI algorithms that feed from data. Exploring how the outcome of AI data-driven algorithms affects explanations is critical for responsible decision-making. These examples show some contextual factors of Big Data Analytics that affect asymmetrical relationships, however, there are other challenges that matter. Next, I will describe some other challenges related to data and AI in more detail.

Challenges for Decision-making

Data-related challenges

The development of digital technologies comes with the inevitable proliferation of data. Digital technologies open the possibility of monitoring all sorts of variables. The working

assumption of Big Data is that extractable insights and patterns buried in data can inform or support decisions. Questioning the act of data collection does not seem to be as relevant as how to extract value from large amounts of data. Then, the pressure to obtain commercial gain from Big Data by identifying valuable patterns builds up (Floridi, 2012). The more pressure, effort, and resources devoted to data exploitation, the more natural it is to think within the boundaries of data. No matter what the problem is, the solution is a matter of digging into the data to discover hidden patterns.

To some extent, it seems everything is in the data. Framing the world through Big Data is problematic for two reasons. First, it reduces the physical world to specific and objective data, disregarding important information relevant to decision-making (aka dataism). Second, meaningless patterns in large data sets may seem meaningful and valid when they are not.

Big Data makes data may look like an objective source of information for us to make sense of. However, data is not objective[4]. Data is created, processed, analyzed, and presented. It goes through multiple processes and translations to determine features that describe a phenomenon. These processes are tinted with human subjectivity, with interpretation, personal interests, perception, and judgment. This means the issue is steered toward specific responses by identifying certain features pertinent to describing a phenomenon. This is problematic because data can be misleading or misrepresented. Moreover, we must consider other aspects of our world that cannot be transformed into data or quantified and are relevant precisely because they cannot be quantified. Focusing excessively on data takes attention from non-quantifiable information that may be relevant for decision-making.

In large datasets, we can see patterns that do not exist because of the number of possible connections (boyd & Crawford, 2012). For example, Tyler Vigen's website "Spurious Correlations" shows patterns between variables where two trends practically fall on top of each other. (Vigen, n.d.). While these are extreme and outrageous cases -like the correlation between the US spending on science and suicides by hanging-they show how numbers can

[4] For more on objectivity and Big Data see: Daston & Galison (2010); Lukoianova (2014); Porter (1988)

be misleading. Data scientists called this “data dredging” (Bergstrom & West, 2020). For Bergstrom & West (2020), correlations “doesn’t mean that there is any meaningful connection between the two trends” (p. 70). Pressure to exploit data exacerbates the dangers of moving too fast from a correlation to a causal connection. Bergstrom & West (2020) show multiple ways in which errors can appear in data—for example, counting mistakes, non-representative or small samples, and faulty procedures stemming from false information. As they put it, numbers can be “fabricated out of whole cloth in an effort to confer credibility on another wise flimsy argument” (p. 81). My point is that data and numbers can be misrepresented to mislead, seeming very convincing and credible.

Machine Learning-Related Challenges

Using Machine Learning algorithms to analyze Big Data leads to concerns about “inconclusive evidence leading to unjustified action and inscrutable evidence leading to opacity” (Mittelstadt et al., 2016; Tsamados et al., 2022). I have discussed how correlation in Big Data can be misleading but seem valid. The probabilistic nature of algorithms is misunderstood by considering associations and correlations as causal connections. Therefore, calling this meaningful information. Additionally, explanations of the connection between inputs and outputs are complicated to achieve. They cannot be generated rapidly because of (1) the multiple possible correlations in large data sets and (2) the complexity of the computational processes and logic proper of Machine Learning algorithms. The lack of explanation hinders the decision-maker's individual moral responsibility.

The fact that many possible correlations can be drawn from data makes it inconclusive. The problem with inconclusive evidence is that it may lead to unjustified actions, therefore becoming a problem for responsibility (Tsamados et al., 2022). As I have discussed, data can correlate erroneously, leading to evidence that may be wrongly used to support decisions. This is not to say that relevant correlations cannot come from large datasets processed by Machine Learning algorithms. However, there is no intrinsic meaning in those patterns. Machine Learning algorithms will produce any patterns. Unfortunately, irrelevant or wrong correlations are sometimes tricky to uncover.

Moreover, even when there is a causal connection, focusing on data may neglect relevant information for decision-making. For example, algorithms that predict patient outcomes rely on quantifiable data ignoring other information that has an impact on the patient, such as their disposition to live (Tsamados et al., 2022, p 218). This shows how the output of data-driven algorithms might not be enough to justify decisions. Non-quantifiable information is important for decision-making, as well as understanding how the relationship between patient-physician is affected in practice by introducing AI algorithms. External factors also affect the decision of the physician, who is the decision-maker.

Concerns around inscrutable evidence go beyond the lack of transparency of Machine Learning algorithms to "the socio-technical infrastructure in which they exist; and the decisions they support" (Tsamados et al., 2022, p. 218). It is not only about the technical complexity of Machine Learning algorithms' outputs but how they are implemented, interpreted, and managed in practical settings. Tsamados et al. (2022) claim that for designers and developers of these algorithms, "lack of transparency often translated into a lack of accountability and lead to a lack of 'trustworthiness'" (p. 218). In short, inconclusive and inscrutable evidence poses a problem that extends the attribution of responsibility for designers and developers.

This section highlighted how Big Data and AI approaches are problematic to solve social and political issues that are relevant for social stability. Trying to address social issues thinking that their complexity can be encapsulated in an algorithm is problematic because it neglects important information for decision making. This neglection exacerbates the problem of providing relevant explanations in decision-making process, which becomes a problem for responsibility. Now, to better understand how Big Data and AI challenge responsibility, we need to discuss in more detail how responsibility is addressed in relation to these technologies. The next section will explore the current debate about responsibility in the light of the challenges posed by AI.

Section 2: Moral Responsibility: Challenges in the Age of AI

This section will explore how responsibility is addressed in the field of ethics of AI in light of the challenges Big Data and AI pose, aiming to problematize individual moral responsibility regarding decision-making by asking questions about how and what the agent is responsible for. This means putting aside questions about the attribution of responsibility and exploring critical aspects to provide a relevant and meaningful explanation to the moral patient.

First, I will begin the debate surrounding "responsibility gaps" and questions about attribution of responsibility. I argue that taking a forward-looking approach that focuses on questions beyond the attribution of responsibility might allow us to think about responsibility differently, providing some insights to mitigate the challenges posed by Big Data and AI. Furthermore, I highlight the relational and interpretative aspects of a forward-looking approach to responsibility as relevant for explanations.

The literature on the notion of responsibility is vast and covers a large number of sub-topics. It can be approached from different perspectives, and several authors have provided taxonomies of responsibility[5]. My purpose is not to provide a comprehensive analysis of the concept or design a checklist to guarantee responsible decisions. This does not mean that it is in vain to discuss responsibility at an abstract level; quite the contrary. A more in-depth understanding and nuances of what responsibility means and our sense of responsibility in different circumstances for different people will bring us closer to more responsible decisions. However, I will start and focus on the current debate about responsibility in AI to better understand what notions of responsibility are challenged.

Discussions about the influence of AI on responsibility are not new. The notion of "responsibility gap," introduced by Andreas Matthias in 2004, has become relevant in the philosophical debate about responsibility with AI. This notion discusses the difficulty of attributing moral responsibility to the manufacturer of "learning machines" for the

[5] For more on responsibility see: Baumgärtner et al. (2018); Davis (2012); Fischer & Ravissa (1998): Hart (1968); Strawson (1962); Van De Poel (2011); Vincent et al. (2011); Zimmerman (1988).

consequences of their operation (Matthias, 2004). More recently, Santoni de Sio & Mecacci (2021) propose four intertwined problems around responsibility gaps "– gaps in culpability, moral and public accountability, active responsibility—caused by different sources, some technical, other organizational, legal, ethical, and societal" (p. 1). Responsibility gaps address the question of who is responsible for certain actions and consequences and therefore are one of the issues related to the control condition for attribution of responsibility[6].

As Matthias (2004) puts it, to attribute responsibility to someone for the consequences of an action, that person should be able to "offer an explanation of her intentions and beliefs when asked to do so" or deserve specific reactive attitudes[7], such as resentment, blame, or praise (p. 175). To deserve these reactive attitudes, the person should meet control and knowledge conditions for responsibility. This means the person should know about the consequences of their action and be in control of freely deciding on a course of action. In this case, being able to provide an explanation is a way to determine if the knowledge condition is fulfilled in order to attribute responsibility to someone for the consequences of certain actions.

The debate surrounding responsibility focuses on questions about attribution of responsibility, in other words, questions about who is responsible for the consequences of certain actions of machines or when machines take over a human task (entirely or partially). As Santoni de Sio & Mecacci (2021) point out, the field ethics of AI has been discussing "to what extent persons can or should maintain responsibility for the behaviour of AI" (p. 1058).

In the same vein, Sven Nyholm, in a seminar given to the Schwarts Reisman Institute about the ethics of AI, provides a matrix with four distinctions about responsibility gaps: negative versus positive and backward-looking versus forward-looking (2023, 27:30).

Nyholm explains that negative responsibility is when someone is to blame for harm or negative consequences that have happened, while positive responsibility is when something

---

[6] For an overview of other issues regarding the control condition, see Coeckelbergh (2020).

[7] For more on reactive attitudes see: Babushkina (2020); Strawson (1962) ; Wallace (2022)

good has happened and someone is to be praised for the positive consequences. Additionally, he explains that when something (positive or negative) has happened in the past, we are talking about backward-looking responsibility, but when we are looking at the future to decide who will be responsible for making sure that certain consequences can be achieved or avoided, we are talking about forward-looking responsibility.

The matrix is about the attribution of responsibility, and Nyholm places different authors in different sections, stating that the forward-looking-positive section is the least explored in the literature. In the same vein, Van de Poel et al. (2015) state, "Most of the philosophical literature on responsibility tends to focus on backward-looking responsibility and often understands backward-looking responsibility in terms of reactive attitudes"(p. 15).

Table 1. Kinds of responsibility gaps (Schwartz Reisman Institute, 2023)

Responsibility Gaps	Backward-looking	Forward-looking
Negative	Most commonly discussed (e.g. Sparrow (2007))	Santoni de Sio & Mecacci (2021)
Positive	Danaher & Nyholm (2021) "Achievement Gaps"	Least discussed. AI value alignment problem?

Even though the distinctions provided in Table 1 are dealing with the question of attribution of responsibility, they are still useful to position this work in the current debate.

Regarding attribution of responsibility, I believe that full control of AI development, deployment, and use should remain with human agents, agreeing with Coeckelbergh (2020) when he says that: "With regard to the two Aristotelian conditions, it is thus assumed that it does not make sense to demand that the AI agent act voluntarily or without ignorance, since an AI agent lacks the preconditions for this" (p. 2054). Therefore, I am not engaging with questions about attribution of responsibility. I am considering only asymmetrical relationships where the user, as the decision-maker, can be held responsible. This is the case

in legal responsibility when in corporations or organizations can be traced to individuals (Coeckelbergh, 2020).

I am aware that the assumption that only humans can be considered responsible agents has some limitations, such as focusing on individual responsibility and the moral agent (Coeckelbergh, 2020) and that this does not solve the problems regarding attribution of responsibility. An agent might not be attributed complete responsibility for the consequences of certain actions, arguing that Aristotelian conditions for responsibility (control and knowledge) were not fully met. Additionally, attribution of responsibility can come in many degrees and depend on different factors and has been problematic even with other technologies before the introduction of AI. For example, in the problem of many hands[8], where many individuals, institutions, and technical artifacts interact with each other, it might be difficult to define who is responsible, making it hard to trace the chain of events and factors leading to a given outcome and its use in decision-making. Or what Nickel et al. (2022) calls "moral uncertainty": uncertainty about how to act or what is the right thing to do as a way in which technology can disrupt society.

In this context, I am positioning myself in the "forward-looking-positive" responsibility section from Table 1 to consider questions about what is the moral agent responsible for and how the agent is going to proceed to achieve that. Stepping aside for a moment from the question of who is responsible and taking a forward-looking approach might help us think about responsibility differnetly. Before moving on, I will elaborate further on the notion of forward-looking responsibility and its connection to backward-looking responsibility.

Backward-looking and Forward-looking responsibility

As mentioned before, I propose a decision-maker as the responsible moral agent, therefore addressing the question of who is responsible. However, as I will show, it is common to address backward and forward-looking approaches in relation to the attribution of

[8] For more on the problem of many hands, see Van De Poel et al. (2015) and (Coeckelbergh, 2020)

responsibility. In what follows, I will elaborate on a forward-looking approach that will consider a moral agent responsible for providing relevant and meaningful explanations.

Retrospective or backward-looking approaches to responsibility look back in time to determine who is responsible for the consequences of actions and the causal chains that led to those actions. In contrast, prospective responsibility takes a forward-looking approach focusing on decisions more than actions and consequences. While both perspectives are connected and necessary, I argue that a prospective or forward-looking approach allows us to consider questions about what the agent is responsible for and how by drawing attention to present decisions that will promote a future state of affairs.

Retrospective approaches stem from legal responsibility, where determining accountability and liability for an action depends on a legal system, its operation, and its enforcement. Legal responsibility is based on a given set of laws and processes established to determine who is liable and for what in a given situation. Legal and retrospective approaches motivate through punishment and social shame. On the other hand, moral responsibility is connected to an obligation to behave according to what is morally right and is related to a forward-looking approach to proactively taking responsibility.

To clarify, imagine the situation where person A is sitting in a public park reading a book. A few meters away at another bench, person B stands up rapidly, leaving a wallet behind. Person A finds the wallet; inside it, there is a significant amount of cash and identification. Legal responsibility pertains to abiding by the laws and regulations. Legally, person A, who found the wallet, would be required to hand it over to the nearest police station. Not doing so could be considered theft and potentially have legal consequences. On the other hand, moral responsibility extends beyond legal obligations to ethical considerations. In this case, the person has a moral responsibility to make reasonable efforts to return the wallet, even if it is not legally required. This could involve trying to identify and get in contact with the owner through the identification cards.

What matters from a legal point of view is accountability and liability. It operates through fear and punishment. Moral responsibility goes beyond legal in the sense that there are no

detrimental consequences for not fulfilling obligations. For moral obligation, motivation comes from virtue and cooperation with others. With a prospective approach, we recognize the impact our choices have on others. Individuals often consider the responsibilities they are willing to take beforehand. When taking a forward-looking approach, we try to understand the other person's position and how our choice might affect them. This means acting compassionately and honestly and being aware of the position of power we are in.

Van de Poel (2011) describes how forward and backward-looking responsibility are connected. In his description of responsibility, there are nine notions, two of which are forward-looking: *responsibility-as-moral obligation*, "as the obligation to see to it that something is the case," and *responsibility-as-virtue*, related to the character or personality of an individual (p. 40). These two forward-looking notions are connected to *responsibility as accountability* and *responsibility as blameworthiness*. The connection implies that an agent who considers and commits to looking forward to a particular state of affairs in the future will be more aware of the accountability and blameworthiness that may derive from not arriving at the aforementioned state of affairs. My point is that forward and backward-looking responsibility are connected, and focusing on forward-looking approaches could serve as a common ground for both perspectives.

Following van de Poel (2011), forward-looking responsibility is connected to responsibility-as-moral obligation and responsibility-as-virtue. This means that responsibility-as-virtue implies that a person voluntarily evaluates and assumes certain responsibilities-as-obligation (van de Poel et al., 2015). This is consistent with Richardson's (1999) two basic components of forward-looking responsibility. First, a disposition to a particular range of concerns and, second, the autonomy to interpret certain rules in order to satisfy those concerns. The first component could be understood as a notion of responsibility-as-moral obligation, while the second one introduces interpretability. For instance, if law enforcement officer P is deployed in a high-risk area, P is responsible for patrolling the area and securing the residents of that area. But that rule might be reinterpreted by P if there is a hostage situation happening in a low-risk area. Under these circumstances, P should leave the initial area to support other officers somewhere else. The attitude to reinterpret a situation in

relation to the specific goals and desirable situations is critical for P to fulfill her responsibility. This also means having the capacity to understand what is required in a certain situation and being able to reinterpret and provide justification for a moral judgment that supports a decision.

The interpretability side raised by Richardson (1999) points out at the relevance of contextual information and human interpretation in decision-making and explanations. When making decisions and providing explanations, individuals should consider specific circumstances, complexities, and nuances of a situation. Contextual information provides a broader understanding of the factors at play, which can impact the decision-making process. Human interpretation works with contextual information and incorporates subjective perspectives, allowing for a more comprehensive explanation. Contextual information and human interpretation enable the moral agent to make informed decisions and provide meaningful explanations that consider the specificities of unique situations.

Going back to responsibility-as-moral obligation, moral obligations may come from the role a person occupies in society. Role responsibility refers to the obligations associated with a particular role within a social institution. Assuming specific roles, such as engineer, healthcare professional, or professor, comes with inherent responsibilities. These responsibilities are typically defined by the expectations and norms associated with the role but are not necessarily moral obligations. The moral agent should evaluate the obligations stemming from the role occupied in society, acknowledging that they are justified by customs or socially accepted values (Babushkina, 2019). This means that obligations attached to the role occupied in society are not necessarily morally right.

Approaches to professional responsibility and responsible research and innovation (RRI) often take a forward-looking stance to responsibility (Davis, 2012; Dignum, 2019; Pesch, 2015; Stilgoe et al., 2013; van de Poel & Sand, 2021). This is mainly because professionals and developers of new technologies bear the responsibility for future consequences of their products and services according to their role.

Developers and designers of technologies, such as AI-CAD or predictive policing, have an impact on the role of other professionals and on social systems. However, my main concern is how these systems influence the responsibility of the moral agent (physician or law enforcement officer). These are two distinct roles. On the one hand, the engineers, designers, and developers of AI applications, and on the other, the role of professionals that use the output of AI to make decisions.

Issues with forward-looking responsibilities are often related to the attribution of responsibility when agents are part of complex social and technical systems where conflicting values and goals combine with unclear roles responsibilities. This lack of clarity in complex systems with different stakeholders can impede a forward-looking stance toward responsibility. Pesch (2015) points out how engineers lack clarity regarding their role and moral values to strive for. Subsequently, this lack of clarity hinders responsibility because engineers do not know their role responsibilities and, therefore, cannot morally evaluate them.

In the context of professional responsibility for engineers working with Artificial Intelligence, Santoni de Sio & Mecacci (2021) discuss two issues related to forward-looking responsibility. First, lack of awareness of social and moral obligations toward others, and second, inability or motivation to fulfill these obligations (Santoni de Sio & Mecacci, 2021, p. 1067). According to the authors, these issues arise when AI is introduced because designers of AI are not fully aware of their responsibility to prevent harm deriving from AI.

As far as I understand it, the introduction of AI exacerbates the problem of institutional clarity posed by Pesch (2015), leading to the issues raised by Santoni de Sio & Mecacci (2021). However, they are not particularly attached to AI but could happen with any other disruptive technology. Nevertheless, the two points that Santoni de Sio & Mecacci (2021) raise are relevant to my discussion of forward-looking responsibility because they apply more broadly to users of technology and individuals that are part of a complex network of social and technical systems.

Building on van de Poel (2015) and Richardson (1999), let us think about what the main goal of a forward-looking approach would be. For van de Poel (2015), the aim of responsibility-as-virtue is “due care to others” (p. 42). How can this be translated to the discussion about responsible decision-making in asymmetrical relationships? Fujita (2020) concludes his article on AI computer-aided detection/diagnostics with a remark from the NTU Center for Data Science stating that the “transition to AI support in diagnostic radiology should proceed like the adoption of self-driving cars—slowly and carefully, building trust, and improving systems along the way with a focus on safety” (p. 17). My point is that responsibility-as-virtue points to care and trust in relationships as a main aim. The question that follows would be how to achieve this aim.

One way could be to reflect on the decision-maker's needs to provide a relevant explanation, for example, by improving digital literacy, translating AI technical language, or increasing knowledge about statistical software (Biswas et al., 2023; Kather, 2023). A forward-looking reflection upon these needs as a requirement to provide explanations means considering the moral patient in the decision-making process. The moral patient requires an explanation for decisions that affect her, more so in a position of vulnerability. In asymmetrical relationships, decision-makers need to explain decisions in a way that is relevant to the moral patient.

Explanations are given to patients or citizens who are in a vulnerable position. Those explanations are relevant to the relationship between doctor and patient or police officer and citizen, both individually and socially. From a vulnerable position, explanations might require reasons that rely on more than accuracy and efficiency.

Additionally, both moral agents and moral patients construct meaning around algorithmic output affecting the capacity to provide explanations and understand them. Technology affects interpretation and, in turn, the capacity to provide reasons for decisions. When decisions are based on algorithmic outputs that lack understanding and scrutability, it becomes challenging to provide meaningful explanations for the reasoning behind their choices. Moreover, understanding the rationale behind algorithmic outcomes and their use is important for the moral patient. In asymmetrical relationships, such as those between doctors and patients, or law enforcement officers and citizens, the ability to understand the

reasoning behind decisions is crucial for building trust and promoting responsible decision-making.

I want to make a distinction between the purpose of explanations from a backward-looking and forward-looking perspective. As Matthias (2004) puts it, to attribute responsibility to someone for the consequences of an action, that person should be able to provide an explanation when asked after a decision is taken and when consequences are often undesired. This would be a backward-looking-negative approach. The purpose of the explanation, in this case, is to determine if someone can be attributed responsibility for the consequences of certain actions that happened in the past. In other words, we are trying to answer who is responsible for the consequences. To clarify, this is framed as a problem for the attribution of responsibility because it is impeding the fulfillment of the control or knowledge conditions. However, the lack of awareness of social and moral obligations toward others, and the inability or motivation to fulfill these obligations (Santoni de Sio & Mecacci, 2021, p. 1067), are problematic beyond the attribution of responsibility because they impede the ability of the agent or user to provide explanations before a decision is taken. Considering explanation as part of the decision-making process would be a forward-looking approach.

So far, I have tried to sketch a notion of forward-looking responsibility building on van de Poel (2015) and Richardson (1999) (as-virtue and as-moral obligation) to explore the relevance of explanations in themselves. This means considering a moral agent focused primarily on the responsibility of providing explanations for decisions. This notion of forward-looking responsibility highlights: (1) a desired state of affairs (forward-looking) where relationships are based on trust and reliance and (2) focused on providing relevant and meaningful explanations (allowing for interpretability and contextual information). With this approach, my intention is to provide answers to the questions of who is responsible and for what: the decision-maker is primarily responsible for providing a relevant explanation before making a decision. A forward-looking approach to responsibility is an individual commitment to actively seek reliable and trustworthy relationships by providing relevant and meaningful explanations. This is relevant for asymmetrical relationships

because it considers the moral patient's needs. This opens up questions of what a relevant and meaningful explanation is and for whom that will be addressed in the following subsection.

Beyond Attribution of Responsibility

Coeckelbergh (2020) states that if the moral agent is not able to explain a decision, this is a problem for responsibility for two reasons. First, the moral agent does not know what she is doing (not fulfilling the knowledge condition for responsibility[9]). Second, "the human agent also fails to act responsibly toward the responsibility patient(s) affected by the action or the decision, who can rightfully demand an explanation for that action or decision since they are affected by it" (Coeckelbergh, 2020, p. 2062). Moreover, he argues that the ethics of AI should foster the moral agent's responsibility in these two senses: First, the agent should know what she/he is doing with the AI, and second, the agent should be responsible in the sense of answerable to those affected (or their representatives)" (p. 2062). This begins to address the questions of what a relevant and meaningful explanation is and for whom by introducing the need for an explanation for the moral patient.

Inconclusive evidence coming from data and the difficulty in explaining algorithmic output impacts the capacity and quality of explanations which becomes a problem for moral responsibility even when algorithms do not act or decide autonomously. For decision-makers, not knowing the relations between inputs and outputs makes it more difficult to explain them to others, degrading their explaining capacity and the quality of those explanations.

Coeckelbergh (2020) discusses responsibility in the sense of answerability. Responsibility as answerability touches upon its relational aspect, which is relevant because it suggests two degrees of moral responsibility for the moral agent: first, in terms of understanding, and second in terms of the moral patient's need for a relevant explanation. Decision makers

[9] See Coeckelbergh (2020) for more on control and knowledge as conditions for responsibility and responsibility as answerability.

should have some degree of knowledge about the technology they are using, its meaning, and its possible effects on their decisions because of how it affects the decision-making process. This would seem like a first step to addressing the first degree of moral responsibility. Secondly, the moral patient requires an explanation for decisions that might affect her. In asymmetrical relationships, explanations can reduce power asymmetry. This approach suggests that responsibility happens in the interaction between moral agent and moral patient.

Even though responsibility as answerability concerns questions about the attribution of responsibility, it allows us to think about alternative questions, such as how we are being responsible, towards whom, and for what. This means shifting from attributing responsibility to taking or being responsible towards someone else. Stepping away from the question of who is responsible may allow for different ways of approaching responsibility that can mitigate the challenges posed by Big Data Analytics.

A Notion of Forward-looking Responsibility: Relational and Interpretative

So far, I have discussed a forward-looking approach to responsibility that includes responsibility-as-virtue, responsibility-as-moral obligation, and responsibility-as-answerability, focusing on questions beyond the attribution of responsibility. In what follows, I will expand on two relevant sides relevant for these notions: relational and interpretative. I propose to pay more attention to the relational and interpretative sides of responsibility in the context of decision-making in asymmetrical relationships by taking a forward-looking approach. The relational side of responsibility has to do with responsibility-as-answerability and the interpretative side with responsibility-as-virtue, where the role of meaning and understanding becomes critical.

For responsibility-as-answerability, the relationship between agent and patient is primordial (Coeckelbergh, 2020). For reliable and trustworthy relationships, I argue that explanations are valuable, not only to determine who is responsible. Explanations play an important role when caring for others, both as a virtue and as a moral obligation for being responsible and towards the moral patient. Explanations, therefore, become relevant for the

three notions of responsibility discussed (virtue, moral obligation, and answerability) and for asymmetrical relationships.

Asymmetrical relationships, such as patient-physician, happen in a context of professional responsibility. In the case of healthcare, physicians hold a professional responsibility to their patients, which goes beyond mere reliance, and involve trust placed in them by the patients. There is a component of interpersonal trust that involves complex emotional reactions and expectations and connects reliance to responsibility (Walker, 2006). Responsibility in asymmetrical relationships should recognize the power dynamics and the inherent vulnerability of the patient through trust. The physician, as a professional with dedicated knowledge and expertise, assumes the responsibility of making informed decisions, communicating effectively, and acting in the patient's best interest while considering their preferences. In an asymmetrical relationship where the vulnerability of the moral patient should be acknowledged, responsibility entails trust and reliance.

Relational Side

Working and collaborating with others is something we do daily, and the concept of responsibility is important to hold relationships together. When discussing responsibility, the relationship between moral agent and moral patient should be considered. In this relationship, the awareness of the position of the moral patient, who will be affected by any decisions taken by the agent, creates an obligation for the agent to explain and justify the decisions taken. Focusing on the condition of the moral patient and the requirement for an explanation should be considered a moral obligation. Coeckelbergh (2020) argues that the demand for an explanation that the moral patient is entitled to translate into a moral requirement for the moral agent to provide an explanation of his/her decisions and actions. Examining the decision-making process should make explanations clearer and more transparent. This is because an explanation is not only about justifying a decision after the fact but before, acknowledging the context and the moral agent in the present.

A meaningful explanation should account for the reasoning and other contextual factor influencing the decision. It should give a picture of the moral agent's thought process,

intentions, and relevant factors. Offering such an explanation demonstrates the importance of trust and vulnerability of the moral patient. Additionally, it recognizes that decisions are not made in isolation but are influenced by dynamic factors, such as the moral agent's own beliefs, emotions, experiences, and limitations.

Related to the relational side of responsibility is the issue of a lack of awareness of social and moral obligations (Santoni de Sio & Mecacci 2021; Pesch 2015). Santoni de Sio & Mecacci (2021) use the example of an engineering manager that believes that the product or service provided will bring more comfort or convenience to the users. However, they are not under the obligation to minimize possible negative impacts on the user's well-being or privacy. This shows a lack of awareness of the social and moral obligations individuals have to others. Acknowledging the relational side can help address the social and moral obligations that individuals have towards others in asymmetrical relationships because it recognizes that responsibility is not single-sided but emerges from interconnected relationships. Providing and receiving explanations could promote a better understanding and evaluation of the expectations and obligations involved. By clarifying intentions, perspectives, and concerns, both moral patients and moral agents can develop a shared understanding of their roles, responsibilities, and the broader context in which their relationship operates. Moreover, doing this could help the evaluation of obligations, their relevance, and justification within social institutions. One cannot choose the norms and expectations in a relationship, but one should be able to evaluate and criticize them (Babushkina, 2019, p. 209).

At a personal and institutional level, explanations promote trust. Walker (2006) discuss individual trust within personal interactions and "default trust" when individuals trust a larger network of agents and allow them to perform daily activities with the confidence that others will behave in a certain way. In asymmetrical relationships, both interpersonal and "default trust" are relevant for responsibility. "Default trust emerges from interpersonal trust becoming part of a relational approach to responsibility.

An objection to this active relational approach to taking responsibility could be that the agents are selfish and prioritize their own personal interests. An ideal of a fully aware and committed moral agent—the objector could argue—is naive and impossible to achieve when

most of the attention goes to the consequences of our actions. While I do not neglect the importance of mainly weighing consequences and personal desires, shifting to a relational forward-looking approach frames the discussion differently. The main difference is focusing on the relationship with the other and acknowledging our interdependence. Individuals are motivated and act in a way that considers the needs of others in relation to their own. While one could act based solely on personal interest, first, this would not mean that following those interest ends up in something good, and second, it neglects the emotional component and importance of relationships between people. In other words, the moral patient acquires relevance in relation to the moral agent, where striving for a strong, long-lasting relationship is the main objective.

Moreover, in human relationships and cooperation with others, we play informal roles that entail implicit obligations. Being a friend, colleague, or neighbor are all informal roles to which obligations and expectations are attached. Alfano (n.d.) notes that people tend to accept and commit to those responsibilities to accomplish a joint endeavor. Observing how we commit to implicit responsibilities in those informal roles means that there is something relevant about the relationship with the other we want to preserve and foster.

The relational side considers providing explanations for decisions as a moral requirement from agents to patients and reflects on the justification. Responsible decision-making should provide the possibility to review the decision-making process and the justifications required by moral patients.

Interpretative Side

The interpretative side is connected to responsibility-as-virtue. Van de Poel (2015) states that "responsibility-as-virtue implies a willingness to actively assume certain responsibilities, and it implies initiative and judgment in taking responsibility (p. 35). My point is that interpretation plays a role in the agent's capacity for judging moral obligations. The pressure exerted by authority or the environment to use technology may impede the moral agent from fulfilling their moral obligations (Santoni de Sio & Mecacci, 2021, p. 1068). The contention of the meaning of technology and how to use it can influence the agent's

capability to judge and fulfill her/his moral obligations. Recall the example where the pressure manifests by trying to impose a meaning of AI-based weapon systems and, subsequently, an understanding of the situation through this technology. An implicit struggle for meaning influences how the user relates to it, understands it, and relates to others affected by it. This situation extends to individuals that relate to technology more broadly and brings into question the meaning and its implications for responsibility.

Grunwald (2020) discusses the relevance of a hermeneutical component in Technology Assessment (TA). Hermeneutics means interpretation and began with the interpretation of religious texts, where different interpretative theories have been developed (Schmidt, 2006). I am comparing TA with responsible decision-making because they both deal with uncertainty and focus excessively on consequences. TA investigates the possible impact of new technologies to guide policy-making, provide information, and contribute to shaping understanding. The emergence of technologies such as nanotechnology and human enhancement represents a challenge for approaches that only evaluate consequences because it is difficult to provide conclusive evidence to support or neglect future scenarios. The increasing uncertainty about future consequences regarding new and emerging sciences and technologies (NEST) is a reason to analyze imaginaries, visions, and expectations that create meanings and trajectories of emerging technology. Building on this concept, TA would focus more on the constitution, challenge, and contention of societal meaning, imaginaries, expectations, and narratives assigned to new technology. Any endeavor that deals with future scenarios must accept a level of uncertainty and needs plasticity to interpret contextual situations and understand multiple perspectives and moral judgments.

Tollon (2020) argues that designers and developers ought to take a hermeneutical analysis. This would involve understanding the context, how stakeholders interact with and understand AI, and how it might affect other stakeholders (Tollon, 2022). Beyond designers, developers, and deployers of technologies, users could also benefit from undertaking a hermeneutical perspective.

Additionally, for decision-makers that rely on technology for decision support, the meaning of technology becomes crucial. They have the responsibility of reflecting on the societal and

personal meaning of the technology used as decision support. Creating awareness about how the agent understands the technology and how it affects decision-making becomes key to responsibility.

Following this line of thought, two levels of understanding can be distinguished. First, understanding relevant to the moral agent. This refers to the process of decision-making, the reasons for an explanation, the problem, the context, and the technology. How is this technology supporting the decision, and why? Second, understanding in relation to the moral patient. This means understanding other perspectives and exploring ways of conveying meaning in decisions. There are multiple ways of explaining something, and not all of them are possible for the moral agent or relevant for the moral patient. It requires effort to tune the possible explanations to improve the relationship. Facilitating understanding and improving clarity at both levels should reduce power imbalance in the relationship between moral agent and moral patient.

The point of looking at decision-making from a hermeneutical perspective is to increase awareness of the impact of meaning on individual choices and a sense of responsibility. One should try to avoid finding oneself, through a decision by another person (related to things that matter, such as health and life), in an undesirable situation where it is unclear how and why one got there, trying to mitigate harm and attributing blame or accountability. A hermeneutical perspective helps to make explicit the role of meaning and understanding in responsible decision-making.

This section brought to light two sides of a forward-looking approach to responsibility. A forward-looking approach to responsibility is an individual commitment to actively seek reliable and trustworthy relationships by providing relevant and meaningful explanations. This is relevant for asymmetrical relationships because it considers the moral patient's needs. It builds on the notions of responsibility as virtue, as moral obligation, and as answerability, arguing for a forward-looking approach to responsibility that goes beyond backward-looking approaches and has two sides worth discussing: relational and interpretative. Furthermore, it discusses the importance of recognizing and integrating these two sides of responsibility to address the impact of Big Data Analytics on asymmetrical

relationships. The two sides, relational and interpretative, highlight two often overlooked elements: interconnectedness and understanding. Acknowledging and emphasizing these aspects of responsibility aims to maintain trust in asymmetrical relationships, prioritize the value of explanations for the moral patient, and enhance the sense of responsibility for the moral agent. Focusing on explanation demonstrates the importance of trust and vulnerability of the moral patient. Additionally, it recognizes that decisions are not made in isolation but are influenced by dynamic factors, such as the moral agent's own beliefs, emotions, experiences, and limitations.

According to van de Poel (2015), the aim of responsibility as virtue is "due care to others" (p. 42). In this vein, the following section explores how to achieve this by elaborating on the importance of explanations and describing a way of caring by explaining. The intention is to describe how a practice of caring through explanations would look by using an analogy to industrial maintenance practices.

Section 3: Responsibility as Maintenance and Care

In this section, I will refer to my experience as an engineer and draw on it to suggest that care, as a goal of responsibility as virtue could be practiced as caring by explaining. By drawing upon the analogy with industrial maintenance, responsibility is practiced as the ongoing effort to provide meaningful explanations, build trust, and nurture relationships. I approach the concept of care through industrial maintenance practices to further elaborate on the relational nature of responsibility, emphasizing how explanations play a role in sustaining relationships and trust.

According to Fahlquist (2015), theories of responsibility-as-virtue[10] have four characteristics: "Responsibility (1) is forward-looking, (2) focused on the person and her relations to other people and the world as opposed to individual actions, (3) requires that the person sees herself as part of a greater context within which she acts, and (4) requires the agent to act in a certain way over time"(p. 192). Furthermore, Fahlquist (2015) argues

[10] For more on responsibility-as-virtue see Young (2006) and Fredriksen (2005).

that “care, moral imagination, and practical wisdom” are the most important “ingredients” of responsibility-as-virtue (p. 192). This is in line with what was said in the previous section and combined with the aim of “due to care to others” (van de Poel et al., 2015, p. 42), connects responsibility-as-virtue to care. The purpose of this section is to elaborate on the fourth characteristic focusing on care, namely the way in which the agent should act, with in asymmetrical relationships in responsible decision-making in. The suggestion is that care could be practiced as caring by explaining.

The maintenance analogy intends to show how caring in maintenance could be translated to decision-making in asymmetrical relationships. This adds a practical way of acting to operationalize care by explaining. The connection between responsibility and maintenance comes from the practice of explaining decisions in relevant terms, which could be considered a practice of care and helps to maintain trusting relationships.

The Concept of Care

The concept of care stems from feminist theory, and it emphasizes the importance of whom we care about when undertaking moral judgments. Care is related to vulnerability, dependence, and connection with others involving support, protection, and commitment. (Shafer-Landau, 2018, p. 281). It acknowledges humans' dynamism, unpredictability, and diversity of individual circumstances accepting that circumstances can change and evolve over time. It recognizes that everyone is unique and influenced by their own set of experiences, desires, and preferences.

In a Different Voice: Psychological Theory and Women’s Development, Carol Gilligan argues that women have a different way of arriving at moral judgment. In her book of 1982, the ethics of care emphasizes the practice of caring for people with whom we have a close relationship, such as family and friends. However, more recent theories have expanded on the concept of care in different areas beyond the idea of women caring for children or family. As Fahlquist(2015) puts it ,“ the most recent theories are applied to wider contexts and show how care can be seen as the central notion for ethics generally. The ethics of care is now not

merely seen as a feminine kind of ethics, but a theory that covers both men and women and that can be applied to most areas" (p. 193).

Care can be considered an emotion, an attitude, or an action (Fahlquist, 2015). Care is about connection to others. "Care is the way in which the world takes significance for us in relation to our interpretation of our interests and the future horizon they foreshadow for us, an activity that unites our emotional, imaginative, and rational sides" (Adam & Groves, 2011). Care also has a forward-looking component, considering possible futures, aspirations, and desires. Moreover, it embraces the emotional, rational, and relational dimensions of human existence and is an effort to understand divergent views and foster responsiveness.

Care is a complex concept that involves various dimensions of the human condition. It embraces relationships, vulnerability, unpredictability, and emotions. It recognizes individuals' diverse experiences and promotes understanding and connection with ourselves and others.

Responsibility is inherent in care and manifested through practical activities within a network of relationships. From a care perspective, responsibility means acknowledging the link between responsibility, reliance, and trust (Walker, 2006). It is not abstract, but it is expressed through specific behaviors and decisions that try to understand others' subjective experiences and explain our own. It can manifest through physical interactions with others and is influenced by our own experiences and the ones of those around us. Therefore, it is not static but a continuous process that requires attention and adaptability.

Maintenance practices are often not as popular as innovation. However, these practices are valuable to maintain stability and social cohesion. In Walker's (2006) terms, these practices are relevant to promote "interpersonal" and "default trust". Looking at responsible behavior in maintenance practices might provide some insights about how to act responsibly in decision-making with AI. In my experience as reliability engineer maintenance practices are creative, caring. I will elaborate on how through explanations maintenance can be a caring practice that fosters reliability, trust, and cohesion. This should strengthen my point about

the importance of explanations in themselves and provide a practical context in which this happens.

Responsibility in Maintenance

Looking at established social practices around power plant maintenance and old technology might give some insights into how relationships are nurtured through maintenance and the relevance of explanations for social structure and trust. In this sense, the practices around maintenance can be considered practices of care.

Industrial maintenance is a practice of care. We might not care about machines in the same sense that we care for human beings, but we still maintain them. Through machines, we relate to each other; we affect each other's lives. Maintaining machines means maintaining relationships between human beings. We can maintain poor or unhealthy relationships or poorly maintain relationships until they break beyond repair. Machines such as rotating machinery are used extensively in industry. For instance, power plants have used pumps, compressors, and turbines for power generation for the past century. Generating electricity is a resource-intensive endeavor that connects people in many ways through machines. Thousands of people work around the clock to ensure energy production. Any decision that an engineer makes affects the work of others: operators, engineers, technicians, sales managers, financial analysts, and users. But also the other way around. Any decision that an operator makes affects others in multiple ways. We relate to each other at different levels and degrees through these machines that operate without interruption. In this realm, planning and reliability are of utmost importance. Maintenance strives to keep things running smoothly, maintaining cohesion and stability. Dhingra & Velmurugan (2015) define maintenance using ISO 14224 from 2006 as: "the combination of technical and associated administrative actions intended to retain an item or system in, or restore it to, a state in which it can perform its required function (p. 1625).

There are three main approaches to maintenance: reactive (aka corrective), preventive, and condition-based monitoring (aka predictive)[11]. Reactive maintenance waits for machines to break to intervene. It is costly and work-intensive. It is always running behind and stressful. Preventive maintenance works with time frames where different jobs are scheduled and done without waiting for something to break. Lastly, condition-based monitoring, also known as predictive maintenance, looks closely at the state in which the equipment operates and tries to identify deviations from regular operating parameters that might indicate a possible failure or malfunction. However, it is not possible to predict the future. Power plant personnel can only measure, process, and analyze a finite number of variables.

Machinery diagnostic engineers are maintainers mostly concerned with condition-based monitoring and root-cause analysis. Condition-based monitoring tries to maintain machinery in optimal operating conditions. It happens continuously in the present, relying on diagnostic tools, experience, interpretation, and communication with other people. It also relies on theories, assumptions, models, and simplifications. In practice, part of the work is understanding the limitations and blind spots of those models and simplifications. Another part is providing others with relevant reasons and explanations for operating parameters and maintenance work decisions. In this sense, machinery diagnostics takes a more comprehensive approach. Similar to medical diagnosis, where the patient can be examined in different ways, but none of them can ever grasp the situation's complexity entirely and explain everything, machinery diagnostics measure and analyze variables to draw relevant conclusions that others can explain and question.

Root cause analysis happens after the fact, while condition-based monitoring constantly seeks to maintain a desired state of affairs. Root case analysis explores the reasons for a malfunction and creates plausible explanations. In this sense, root cause analysis takes a backward-looking approach. It looks at who to blame and how. Doing root cause analysis

[11] For more on maintenance approaches and strategies, see: Dhingra & Velmurugan (2015); Gackowiec(2019); Patil et al. (2021).

means asking difficult questions, pointing to inconsistencies and a lack of reasons for actions and decisions.

Condition-based monitoring continuously measures and monitors deviations from the operating parameters. This implies defining the limits or thresholds of those parameters to identify deviations. When defining those thresholds, explanations and reasons that justify those definitions are necessary. Understanding and trusting the underpinnings of those definitions allows us to base decisions on them. The definition of such thresholds may have come from heuristics, rules of thumb, or reasonable approximations, but that could be questioned in the future.

Any deviation or variation is considered meaningful if the threshold of a parameter is explained, justified, and understood. It is an indication of normal operating conditions agreed between engineers and operators. Working under these normal operating conditions is necessary, but that does not mean that those normal levels cannot be questioned. However, to question responsibly established thresholds, there is a need for reasons that back up concerns or inquiries about the thresholds. This means that responsible threshold definition requires explanations, justifications, and understanding among people. It is a process that happens continuously among engineers and operators.

Irresponsibility in threshold definition means that those thresholds are arbitrarily chosen or believed to be arbitrarily chosen, creating uncertainty, unreliability, and distrust. This means that the explanations and reasons to establish those thresholds are ungraspable or unavailable (physically or conceptually). One could argue that some people do not want to understand or are not interested in explanations. While plausible, this does not warrant the lack of relevant, clear reasons and explanations.

Having clearly established and meaningful operating parameters and thresholds create a feeling of trust in a power plant, and people are confident in making decisions based on deviations. Any further decisions based on them are considered valid and legitimate. This is consistent with default trust (Walker, 2006), which emerges when individuals trust institutions. In this case, individuals trust a larger network of agents and allows them to

perform daily activities with the confidence that others will behave in a certain way. Default trust is linked to interpersonal trust. In the context of power plant operation, the operator relies on the condition-monitoring engineer to provide relevant and meaningful operating parameters. This means that responsibility involves more than mere reliance. The engineer is reliable in the sense that she behaves accordingly to what is expected. However, trust happens in relation to the operator that trusts the engineer not only to behave as expected but to do it considering the operator's understanding.

Expectations come from the role a person occupies and from interpersonal relations. Understanding this relational aspect of trust also means paying attention to the way how certain normative expectations are fulfilled. It also means understanding that for the operator -moral patient-, it is not only about reliance but trust, which is an emotional and relational concept. Taking responsibility for a decision involves both the engineer and the operator. As (Walker, 2006) points out, trusting relations are extremely complex, and so are responsibility relations. Developing traceable, explainable, and understandable parameters and thresholds gives others a feeling of reliability and trustworthiness, making their decision-making process and sense of responsibility much higher. This means that questioning those thresholds requires an understanding of the actual situation and being able to trace the process.

The success of root cause analysis and condition-based monitoring depends on people and how they relate to each other. The success or failure of both depends on communication and responsibility. It depends on what kind, how, and why we maintain relationships. Regardless of the most accurate and precise tools or the most expensive and prestigious machines, people matter the most.

While condition-based monitoring is the most promising on paper, it is also the most complex to practice because it depends precisely on social practices among people and institutions. No technology can be successfully implemented where human relations do not work. While here we are merely talking about a machine's "health," it underpins a power plant's social structure and functioning, a social structure where people rely on and relate to

each other. Relationships, for better or worse, are constructed through practices of maintaining machines.

What happens when we strip the reasons and explanations from the condition-monitoring and root cause analysis? Interpersonal responsibility weakens, and so does reliability and trust. Low levels of default trust are negative because they generate negative attitudes toward others and obstruct the development of normal activities, as Walker (2006) points out. Moreover, we cannot question ourselves and others when we cannot provide meaningful explanations and reasons. What does condition monitoring work mean if we cannot explain it to others? Or if we try to explain it to someone that is not interested in understanding?

While the maintenance description of a power plant may seem unrelated to the discussion of responsibility and social structures, it is an illustrative example highlighting three aspects. Firstly, the complexity and uncertainty of engineering systems, such as a power plant, demonstrate the limitations of our ability to control and understand these systems fully. However, we can still make sense of them by developing processes, strategies, practices, and knowledge to ensure their proper functioning and maintenance. Recognizing complexity and uncertainty and how to work around them to make sense of certain aspects of the world is important for responsibility.

Second, the description shows the significance of understanding the social dynamics in which technology operates. A power plant is not simply a technical apparatus; it is embedded within a larger social context involving various stakeholders, including operators, engineers, regulators, governmental institutions, and society. Recognizing and considering the social dimensions helps comprehend the broader impact and implications of technology in decisions.

Lastly, it shows the importance of meaningful reasons and explanations in maintaining and strengthening social structures. Providing clear and understandable explanations for decisions is a way of establishing trust and reliability between moral patient and moral agent. Meaningful explanations foster understanding, address concerns, and justify

maintenance and care practices, contributing to high levels of default trust and cooperation. Being responsible within maintenance involves engaging in practices encompassing understanding and explanations. It requires recognizing the complexity and uncertainty of the system, the social dynamics in which it operates, and the importance of providing meaningful explanations to strengthen "trusting relationships" and increase "default trust" (Walker, 2006).

Conclusion

This thesis aimed to explore the challenges of Big Data and AI for responsible decision-making in asymmetrical relationships. Through a literature review, conceptual analysis, and analysis of practical cases, I argue that Big Data and AI hinder the ability to provide relevant explanations, which is a problem for responsibility. Taking a forward-looking approach to responsibility drawing upon a feminist perspective and incorporating insights from my own experience with industrial maintenance practices, this research has shed light on the importance of providing relevant explanations for asymmetrical relationships in decision-making processes.

Big Data Analytics provides a framework to understand a phenomenon by analyzing vast amounts of data, which can help reduce the uncertainty and complexity of certain situations. However, this reduction often comes with a downside. It oversimplifies the world to data and narrows down our sense-making capabilities to cognitive processes. The main problem is not the simplification but the assumption that this simplification provides a unique and meaningful interpretation of the world. Algorithms are mathematical constructs that struggle that to account for the materiality, ambiguity, and uniqueness of human experience encountered in social and political issues. This is problematic when we try to use their outcome to inform decisions about social issues, such as predictive policing, where the intention is to predict human behavior.

Moral agents and moral patients construct meaning around algorithmic output affecting the capacity to provide explanations and understand them. Technology affects interpretation and, in turn, the capacity to provide reasons for decisions. When trying to address social

problems under an oversimplified framework, the capacity to provide relevant explanations diminishes. The impact of poor explanations has far-reaching consequences for responsibility in asymmetrical relationships. When decision-makers fail to provide meaningful explanations for their decision, it affects responsibility beyond attribution, leading to unreliability and distrust. This weakens relationships and social structures, pushing individuals towards feelings of distrust, anger, and despair.

In the context of AI and Big Data, the introduction of complex algorithms can further complicate the ability to provide meaningful explanations, making obligations and responsibilities appear blurry and difficult to comprehend. Individuals may be uncertain about what they are responsible for and feel disconnected from the consequences of their decision. These issues challenge traditional conditions for attribution of responsibility and are largely discussed in the literature, often from a backward-looking approach where explanations are helpful to determine who is responsible for the consequences of certain actions. However, explanations are critical to fostering and nurturing reliable and trusting relationships during decision-making processes- not only as a way to attribute responsibility after the fact. The lack of them is a major concern because it negatively affects the relationship between moral agent and moral patient.

Moral patients, who are often in a vulnerable position, suffer from this lack of meaningful explanations, perpetuating power imbalances and increasing their vulnerabilities. The capacity and willingness to understand and provide explanations play an important role in promoting or demoting responsibility. A lack of commitment to providing clear and meaningful explanations hinders decision-making and diminishes engagement with these explanations.

This thesis calls for a critical engagement with the challenges of Big Data Analytics on decision-making and responsibility by recognizing the limitations of algorithmic outputs and acknowledging its impact on the capacity to provide explanations. These limitations include framing a phenomenon through data and feature selection, the unquestionable acceptance of algorithmic outputs, and the challenge of understanding algorithmic decision-making processes. It prompts reflection on the role of explanations for responsibility and how social

and contextual factors influence the capacity to provide and understand explanations. The explanations we seek are not related to how or why an algorithm reaches a certain outcome or how strong is correlation/causation between inputs and outputs. Relevant and meaningful explanations consider the need of the moral patient and explain how the algorithmic outcome is used, why, and for what purposes. It also reflects upon the understanding of what the algorithm is and what are the expectations around it. Explanations are given during the decision-making process and their purpose is to connect moral agent with moral patient.

This thesis proposes a forward-looking approach to responsibility based on three notions, responsibility as virtue, moral obligation, and answerability (Coeckelbergh, 2020; Richardson, 1999; Van De Poel, 2011; van de Poel et al., 2015); highlighting relational and interpretative elements and showing how care, trust, and understanding depend on explanations and affect responsibility. This implies a different approach to explanations. Relevant and meaningful explanations happen during the decision-making process between agent and patient. Their purpose is to increase and maintain trust and reliability in asymmetrical relationships. These explanations should consider the limitations of applying algorithms to address social issues, the impact of the meaning of technology in decision-making, the contextual factors involved, and the needs of the moral patient, and should answer why and how the algorithm and its output participate in the process.

To preserve trust and promote responsible decision-making, it is crucial to prioritize explanations that go beyond mere accuracy and efficiency. Decision-makers should make an effort to engage with those affected by their decisions, actively involving moral patients in the process and recognizing their requirements for meaningful explanations. Acting responsibly means caring about others by explaining decisions during the process. An analogy of industrial maintenance practices is used to argue for the importance of meaningful explanations and show how caring by explaining could look like. Maintenance practices show it is fundamental to acknowledge the bi-directionality of relationships between the moral agent and the moral patient, ensuring that both are actively involved in the decision-making process.

There are several limitations in this work that should be noted. Alternative definitions and approaches to AI from other disciplines, such as social sciences, have not been discussed. A constructive criticism of computer science approaches and the narrative they promote could have explored more accurate alternatives.

The literature analysis on the concept of responsibility is limited. It should have been more comprehensive and detailed. Other approaches and notions were not discussed and could shed light on how to address the issues posed by Big Data and AI. Moreover, the concept of forward-looking responsibility could be studied further in the literature in connection to AI and explanations.

Feminist theory as a philosophical framework theory could have been explained and researched in more detail. For example, a more in-depth analysis of the issues through the lens of ethics of care could have been more appropriate. Moreover, decision-making theories and models were not discussed and should have been considered from a philosophical perspective in relation to the role of explanations. Studying decision-making models and theories might provide insights into human decision-making and factors that affect it that are relevant when including algorithms in the process.

Regarding trust and decision-making, mistakes are inevitable. However, I did not elaborate further on how this approach can repair damages caused by wrong decisions. This could have been discussed further, starting with Walker's (2006) concept of "moral repair".

The description of maintenance practices is based on personal experiences and limitations concerning asymmetrical relationships in engineering practice, compared to other domains, could have been explained. Future research could aim at establishing empirically what a relevant and meaningful explanation consists of for different stakeholders. Moreover, other mechanisms that show how AI should support decisions could be explored.

To conclude, this thesis has shown the challenges Big Data Analytics poses for responsible decision-making in asymmetrical relationships in terms of explanations and has explored a forward-looking approach to responsibility. This approach helps us in addressing the

challenges posed by Big Data and AI by focusing on how to achieve relevant and meaningful explanations as a practice of caring for others during decision-making processes.

References

Adam, B., & Groves, C. (2011). Futures Tended: Care and Future-Oriented Responsibility. *Bulletin of Science, Technology & Society*, *31*(1), 17–27. https://doi.org/10.1177/0270467610391237

Babushkina, D. (2019). Bradley's "my station and its duties" and its moral (in)significance. *Zeitschrift Für Ethik Und Moralphilosophie*, *2*(2), 195–211. https://doi.org/10.1007/s42048-019-00049-0

Babushkina, D. (2020). Robots to Blame? In M. Nørskov, J. Seibt, & O. S. Quick (Eds.), *Frontiers in Artificial Intelligence and Applications*. IOS Press. https://doi.org/10.3233/FAIA200927

Baumgartner, S., Petersen, T., & Schiller, J. (2018). The Concept of Responsibility: Norms, Actions and Their Consequences. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3157667

Bell, J. (2022). What Is Machine Learning? In S. Carta (Ed.), *Machine Learning and the City* (1st ed., pp. 207–216). Wiley. https://doi.org/10.1002/9781119815075.ch18

Bergstrom, C. T., & West, J. D. (2020). *Calling bullshit: The art of skepticism in a data-driven world* (First edition). Random House.

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, *2*(2), 100205. https://doi.org/10.1016/j.patter.2021.100205

Biswas, S., Biswas, S., Awal, S. S., & Goyal, H. (2023). Artificial intelligence & deep learning for the radiologist: A simple updated guide without the maths. *Chinese Journal of Academic Radiology*, *6*(1), 7–9. https://doi.org/10.1007/s42058-022-00113-6

Bjerring, J. C., & Busch, J. (2021). Artificial Intelligence and Patient-Centered Decision-Making. *Philosophy & Technology*, *34*(2), 349–371. https://doi.org/10.1007/s13347-019-00391-6

boyd, danah, & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Buhmann, A., Paßmann, J., & Fieseler, C. (2020). Managing Algorithmic Accountability: Balancing Reputational Concerns, Engagement Strategies, and the Potential of Rational Discourse. *Journal of Business Ethics*, *163*(2), 265–280. https://doi.org/10.1007/s10551-019-04226-4

Coeckelbergh, M. (2020). Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability. *Science and Engineering Ethics*, *26*(4), 2051–2068. https://doi.org/10.1007/s11948-019-00146-8

Danaher, J., & Nyholm, S. (2021). Automation, work and the achievement gap. *AI and Ethics*, *1*(3), 227–237. https://doi.org/10.1007/s43681-020-00028-x

Daston, L., & Galison, P. (2010). *Objectivity* (1. paperback ed). Zone Books.

Davis, M. (2012). “Ain’t No One Here But Us Social Forces”: Constructing the Professional Responsibility of Engineers. *Science and Engineering Ethics*, *18*(1), 13–34. https://doi.org/10.1007/s11948-010-9225-3

Dhingra, Dr. T., & Velmurugan, R. (2015). Maintenance Strategy Selection and its Impact on Maintenance Function—A Conceptual Framework. *International Journal of Operations & Production Management*. https://doi.org/10.1108/IJOPM-01-2014-0028

Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer International Publishing. https://doi.org/10.1007/978-3-030-30371-6

Fahlquist, J. N. (2015). Responsibility as a Virtue and the Problem of Many Hands. In *Moral Responsibility and the Problem of Many Hands*. Routledge.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. https://philpapers.org/rec/FISRAC-3

Floridi, L. (2012). Big Data and Their Epistemological Challenge. *Philosophy & Technology*, *25*(4), 435–437. https://doi.org/10.1007/s13347-012-0093-4

Fujita, H. (2020). AI-based computer-aided diagnosis (AI-CAD): The latest review to read first. *Radiological Physics and Technology*, *13*(1), 6–19. https://doi.org/10.1007/s12194-019-00552-4

Gackowiec, P. (2019). General overview of maintenance strategies – concepts and approaches. *Multidisciplinary Aspects of Production Engineering*, *2*, 126–139. https://doi.org/10.2478/mape-2019-0013

Geels, F. W. (2004). From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory. *Research Policy*, *33*(6), 897–920. https://doi.org/10.1016/j.respol.2004.01.015

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

Grindrod, P. (2014). Introduction: The Underpinnings of Analytics. In P. Grindrod (Ed.), *Mathematical Underpinnings of Analytics: Theory and Applications* (p. 0). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198725091.003.0001

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

Hart, H. L. A. (1968). Punishment and Responsibility. *Philosophy*, *45*(172), 162–162.

Jenga, K., Catal, C., & Kar, G. (2023). Machine learning in crime prediction. *Journal of Ambient Intelligence and Humanized Computing*, *14*(3), 2887–2913. https://doi.org/10.1007/s12652-023-04530-y

Kather, J. N. (2023). Artificial intelligence in oncology: Chances and pitfalls. *Journal of Cancer Research and Clinical Oncology*. https://doi.org/10.1007/s00432-023-04666-6

Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. SAGE Publications Ltd. https://doi.org/10.4135/9781473909472

Kwame, A., & Petrucka, P. M. (2021). A literature-based study of patient-centered care and communication in nurse-patient interactions: Barriers, facilitators, and the way forward. *BMC Nursing*, *20*(1), 158. https://doi.org/10.1186/s12912-021-00684-2

Lukoianova, T., & Rubin, V. L. (2014). Veracity Roadmap: Is Big Data Objective, Truthful and Credible? *Advances in Classification Research Online*, *24*(1), 4. https://doi.org/10.7152/acro.v24i1.14671

Maeseneer, J. D., Weel, C. van, Daeren, L., Leyns, C., Decat, P., Boeckxstaens, P., Avonts, D., & Willems, S. (2012). From “patient” to “person” to “people”: The need for integrated, people centered health care. *International Journal of Person Centered Medicine*, *2*(3), Article 3. https://doi.org/10.5750/ijpcm.v2i3.148

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

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*, *6*, 26094. https://doi.org/10.1038/srep26094

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 205395171667967. https://doi.org/10.1177/2053951716679679

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1. https://doi.org/10.1186/s40537-014-0007-7

Nickel, P. J., Franssen, M., & Kroes, P. (2010). Can We Make Sense of the Notion of Trustworthy Technology? *Knowledge, Technology & Policy*, *23*(3), 429–444. https://doi.org/10.1007/s12130-010-9124-6

Nickel, P. J., Kudina, O., & van de Poel, I. (2022). Moral Uncertainty in Technomoral Change: Bridging the Explanatory Gap. *Perspectives on Science*, *30*(2), 260–283. https://doi.org/10.1162/posc_a_00414

Otterlo, M. (2013). A Machine Learning View on Profiling. *Privacy Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*. https://doi.org/10.4324/9780203427644

Patil, A., Soni, G., Prakash, A., & Karwasra, K. (2021). Maintenance strategy selection: A comprehensive review of current paradigms and solution approaches. *International*

Journal of Quality & Reliability Management, *39*(3), 675–703. https://doi.org/10.1108/IJQRM-04-2021-0105

Pesch, U. (2015). Engineers and Active Responsibility. *Science and Engineering Ethics*, *21*(4), 925–939. https://doi.org/10.1007/s11948-014-9571-7

Porter, A. (1988). Indicators: Objective Data or Political Tool? *The Phi Delta Kappan*, *69*(7), 503–508.

Richardson, H. S. (1999). Institutionally Divided Moral Responsibility*. *Social Philosophy and Policy*, *16*(2), 218–249. https://doi.org/10.1017/S0265052500002454

Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson Education.

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Schmidt, L. K. (Ed.). (2006). Introduction: What is hermeneutics? In *Understanding Hermeneutics* (pp. 1–9). Acumen Publishing. https://doi.org/10.1017/UPO9781844653843.001

Schwartz Reisman Institute, S. (Director). (2023). *Sven Nyholm | AI, responsibility gaps, and asymmetries between praise and blame*. https://www.youtube.com/watch?v=T9Er8DkROg0

Shafer-Landau, R. (2018). *The fundamentals of ethics* (Fourth edition). Oxford University Press.

Southerton, C. (2020). Datafication. In L. A. Schintler & C. L. McNeely (Eds.), *Encyclopedia of Big Data* (pp. 1–4). Springer International Publishing. https://doi.org/10.1007/978-3-319-32001-4_332-1

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42*(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Strawson, P. (1962). Freedom and Resentment. *Proceedings of the British Academy*, *48*, 187–211.

Tollon, F. (2022). Is AI a Problem for Forward Looking Moral Responsibility? The Problem Followed by a Solution. In E. Jembere, A. J. Gerber, S. Viriri, & A. Pillay (Eds.), *Artificial Intelligence Research* (pp. 307–318). Springer International Publishing. https://doi.org/10.1007/978-3-030-95070-5_20

Tsamados, A., Aggarwal, N., Cowls, J., Morley, J., Roberts, H., Taddeo, M., & Floridi, L. (2022). The ethics of algorithms: Key problems and solutions. *AI & SOCIETY*, *37*(1), 215–230. https://doi.org/10.1007/s00146-021-01154-8

Turing, A. M. (1950). COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Van De Poel, I. (2011). The Relation Between Forward-Looking and Backward-Looking Responsibility. In N. A. Vincent, I. Van De Poel, & J. Van Den Hoven (Eds.), *Moral Responsibility* (Vol. 27, pp. 37–52). Springer Netherlands. https://doi.org/10.1007/978-94-007-1878-4_3

van de Poel, I., Royakkers, L., & Zwart, S. D. (2015). *Moral Responsibility and the Problem of Many Hands* (0 ed.). Routledge. https://doi.org/10.4324/9781315734217

van de Poel, I., & Sand, M. (2021). Varieties of responsibility: Two problems of responsible innovation. *Synthese*, *198*(19), 4769–4787. https://doi.org/10.1007/s11229-018-01951-7

Vigen, T. (n.d.). *Spurious correlations*. Retrieved June 20, 2023, from http://tylervigen.com/spurious-correlations

Vincent, N. A., van de Poel, I., & van den Hoven, J. (Eds.). (2011). *Moral Responsibility: Beyond Free Will and Determinism* (Vol. 27). Springer Netherlands. https://doi.org/10.1007/978-94-007-1878-4

Walker, M. U. (2006). *Moral Repair: Reconstructing Moral Relations after Wrongdoing*.

Wallace, R. J. (2022). Responsibility and Reactive Attitudes. In D. K. Nelkin & D. Pereboom (Eds.), *The Oxford Handbook of Moral Responsibility* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190679309.013.32

Wang, P. (2008, June 20). *What Do You Mean by "AI"?* Artificial General Intelligence. https://www.semanticscholar.org/paper/What-Do-You-Mean-by-%22AI%22-Wang/2cafacb966ad7ea5e219175b52f4a8d708772c96

Zhou, N., Zhang, C., Lv, H., Hao, C., Li, T., Zhu, J., Zhu, H., Jiang, M., Liu, K., Hou, H., Liu, D., Li, A., Zhang, G., Tian, Z., & Zhang, X. (2019). Concordance Study Between IBM Watson for Oncology and Clinical Practice for Patients with Cancer in China. *The Oncologist*, *24*(6), 812–819. https://doi.org/10.1634/theoncologist.2018-0255

Zimmerman, M. J. (1988). *An essay on moral responsibility*. https://philpapers.org/rec/ZIMAEO