Charlene Palomer Hinton

The State of Ethical AI in Practice: A Multiple Case Study of Estonian Public Service Organizations

Master Thesis

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

Supervisor: Dr. Bettina Distel
Tutor: None

Presented by: Charlene Palomer Hinton (509651)
Schlossplatz 2
48149 Münster
+49 251 8338100
chinton@uni-muenster.de

Date of Submission: 2021-08-09

Content

Figures

Tables

Abbreviations

AI	Artificial Intelligence
EU	European Union
GDPR	General Data Protection Regulation
MIT	Massachusetts Institute of Technology
VSD	Value Sensitive Design

Acknowledgements

Completing this thesis would not have been possible without the generosity of the public servants and partners in the Estonian public administration who have given me the privilege of their time and knowledge in sharing their experiences on AI development. To them I give my sincerest thanks.

I would also like to thank the Erasmus+ PIONEER program for the experience throughout the past two years. The professors and supervisors from KU Leuven, WWU and TalTech have imparted with me lessons and knowledge that I hope to apply as I move onto the next chapter.

To my PIONEERS – 3rd cohort – thank you for the friendship, laughter and memories throughout this journey.

To myself, for fulfilling a lifelong dream of completing a master's program. Charlene, you did it. This is for you.

Abstract

Despite the prolific introduction of ethical frameworks, empirical research on AI ethics in the public sector is limited. This empirical research investigates how the ethics of AI is translated into practice and the challenges of its implementation by public service organizations. Using the Value Sensitive Design as a framework of inquiry, semi-structured interviews are conducted with eight public service organizations across the Estonian government that have piloted or developed an AI solution for delivering a public service. Results show that the practical application of AI ethical principles is indirectly considered and demonstrated in different ways in the design and development of the AI. However, translation of these principles varied according to the maturity of the AI and the public servant's level of awareness, AI knowledge and competences. Data-related challenges persist across as public service organizations work on fine-tuning their AI applications.

1 Introduction

Artificial intelligence (AI) has deep potential to change various aspects of citizen's daily lives and society (Berryhill et al., 2019; van Noordt & Misuraca, 2020). Although a universal definition has not been agreed upon, Wirtz, Weyrer, and Greyer refer to AI as the "capability of a computer system to show human-like intelligent behavior" with core skills including the ability to learn, understand and perceive (2020, p. 599).

A systematic review of academic literature has shown a growing interest in the uptake of artificial intelligence in the public sector (Gomes de Sousa et al., 2019). In Europe, the use of AI in public services is increasing, with over 230 empirical use cases identified (Misuraca, G., & van Noordt, C, 2020). AI applications bring significant benefits to institutions that deploy them, from improving public services to reducing the costs and administrative burden (Mehr, 2017; Misuraca et al., 2020). However, these benefits are countered with sobering risks. Concerns for citizen's privacy and security, loss of decision-making autonomy, and unintentional harm that arise from AI systems that may reinforce existing discriminatory practices (Sun & Medaglia, 2019).

As a response to the risks, international organizations and institutions have increasingly advocated for the ethical design and development of AI. The results of their endeavors are realized through the introduction of ethical guidelines, standards, and governance frameworks, or soft law (Bartneck et al., 2021). More recently concrete actions toward operationalizing ethics have emerged in the form of legislative proposals for AI (EU Proposal AI Regulation, 2021). As technical developments in AI flourish, the ethics of AI persists as a contentious yet important discussion for society, putting into question the human values that are deemed important by society.

Against the background of the expanding, multidisciplinary field of AI, empirical research on AI in the public sector has been inadequate (Sun & Medaglia, 2019; Zuiderwijk et al., 2021). Even less has been published on the practical implementation of ethics of AI in this sector. Only a handful of empirical studies address the state of AI ethics in practice, and they have either focused on companies in the private sector (Vakkuri et al., 2020) or on a broad mix of both (Desouza et al., 2020; Ryan et al., 2021).

Researchers note that in practice, most governments have limited understanding of the implications of the use of AI. They hypothesize that insufficient research on empirical, context-based AI use in governments can induce serious, systemic failures that may negatively impact not only governments but also societies as a whole (Zuiderwijk et al., 2021).

Therefore, the aim of this research to address this knowledge gap in the rapidly-evolving field of AI by addressing the following questions:

How do public service organizations ensure ethically-aligned AI public services in practice?

A. What are the key issues that public service organizations face in AI design and development?

B. In what ways are AI ethical principles considered in practice by public service organizations in the design and development of AI for public service delivery?

By answering these questions, this empirically-grounded, multiple case research contributes to broader academic discussions on the practical implementation of AI ethics and concurrently maintaining focus on the insufficiently researched public sector in the AI discipline.

The rest of this research is organized as follows. Section 2 offers research background on the key concepts and challenges relevant to AI's application in the public sector as well as the debates concerning AI ethics in practice as discussed in literature. Section 3 presents the Value Sensitive Design framework, the theoretical lens through which the research questions are addressed. Section 4 details the methodology used to prime the research analysis and guide the inquiry on the practical translation of AI ethics by public service organizations. Section 5 presents the empirical results that emerged from this analysis, the implications from which are critically discussed in Section 6. Finally, Section 7 concludes with a summary of the findings and future avenues of research.

2 Research Background

2.1 Literature Review Method

Relevant literature to facilitate the understanding of this research are drawn from various academic disciplines including computer science, information systems, law, ethics and public administration. Related research articles are retrieved from reputable databases: ProQuest, Web of Science, Scopus, and the Digital Government Reference Library. Because AI is a multidisciplinary topic, the queries used have been constructed in deference to keywords such as “artificial intelligence,” “public sector,” as well as reference to “ethics.”

Additional keywords are used to further expand the scope of the query, particularly other terms for AI and public sector. For example, machine learning, algorithms, facial recognition, self-driving vehicles, chatbots, and so forth are terms that could be related to AI. Thus, it is necessary to include these terms in order to cast a wide net and obtain the most relevant literature from these academic sources.

Furthermore, since AI is a field that is gaining momentum in the research community, research for which is performed not only by academia but also by governments, think tanks, international organizations, firms and the like. As a result, some timely and appropriate contributions to this body of knowledge are published beyond the formal due process of academic, peer-reviewed journals in the form of white papers. It is, therefore, deemed important to consider these sources, however, interpreting their methods and results critically with as much caution and intellectual skepticism.

The literature review is structured in four main areas: the definition of AI, the state of its use in the public sector, the ethics of AI and AI ethics in practice. The remaining sections discuss the theoretical framework underpinning this research and an introduction to the selected country case.

2.2 Defining AI

Although enthusiasm for AI is on the rise once again, AI is not a new field (Sun & Medaglia, 2019). Research interests in AI have progressed since post-World War II through a machine when several mathematicians formalized answers related to questions and the ability to solve them computationally. In 1950, Alan Turing, a Cambridge mathematician published a paper that hinted at the possibility of artificial intelligence and provided a series of questions to determine the existence of AI, which came to be known as the Turing Test (Franklin, 2014). In the United States, Dartmouth math professor John McCarthy first used the term "artificial intelligence" in 1955 at a conference (Davenport et al., 2019).

Over the next few years, substantive claims were made by intellectuals on the possibilities that AI would bring. Economist Herbert Simon suggested that in a game of chess, computers would beat humans within a decade, although history showed it would take 40 years before this would happen. Along the same tone of optimism, cognitive scientist Marvin Minsky claimed that problems associated with creating AI would be resolved within a generation, however, these challenges exist well beyond a generation (Davenport et al., 2019).

Mainly attributed to the lack of computing technology at the time and the inability to incorporate contextual knowledge, repeated failures of AI application, specifically in machine translations, dampened the enthusiasm for the prospects of AI systems' ability to exhibit intelligence akin to humans (Russell & Norvig, 2021). This skepticism was further nurtured by outspoken critics of machine intelligence of the time.

Former MIT philosophy professor Dreyfus AI opened the debate to the limitations of computers and of AI (1974). His theory argues that AI is based on two assumptions which are false: 1) that intelligence is fundamentally information processing and 2) that knowledge can be transposed into independent, discrete representations (Susser, 2013).

However, key to this argument is the belief that there is a distinction between two facets of human intelligence. One one hand, there is the knowing that, or the factual knowledge and reasoning about it. On the other hand, there is the knowing how, or the skills or behaviors, associated with acquiring or exhibiting this knowledge. The computational view of AI considers first the knowing that as foundational while knowing how becomes a matter of complexity. Dreyfus demonstrated that this view is false as it is backwards because it is our ability to function and adapt to the world around us, or our know-how, that allows us to create factual knowledge (Susser, 2013). At its core, Dreyfus argues that

meaning is inherently context-dependent, and it therefore follows that meaning cannot be formalized into discrete representations for computational programming.

Following this, a lull in AI research occurred in the early 1970s when the limitations of computing as they relate to neural nets and classification became magnified, ushering AI developments into a period called the AI 'winter' which lasted until the late 1990s (Franklin, 2014, Sun & Medaglia, 2019). Dreyfus's line of criticism of the limitations of AI is continued in the early 2000s by Crevier who stated that " no amount of finessing or footwork would ever let a machine do common-sense logic" (1993, p. 240).

Since then, the academic community has wrestled with defining what is meant by 'artificial intelligence' (Etscheid, 2019; Grosz et al., 2016; Legg & Hutter, 2007; Simmons & Chappell, 1988; Wirtz et al., 2020). Early attempts to define artificial intelligence have amounted to anthropomorphic descriptions of systems behaviors such as problem solving and pattern recognition. Simmons and Chappell proposed a definition in which AI denotes behavior of a machine in the same way as a human were to behave, and therefore it can be considered intelligent (1988). The definitions during this time were directly linked to the cognitive abilities of humans, but abilities that can be replicated in machines.

A decade later, other propositions surfaced, with similar linkages to the previous. For example, Barth defined AI as the "pursuit of machine or computer intelligence that approximates the capabilities of the human brain" (Barth & Arnold, 1999, p. 333). Early definitions of AI capabilities therefore suggest human-level intelligence.

It is important to note, however, that the conundrum of defining AI is due in part to the lack of a standard definition for intelligence itself. What is exactly meant by the term intelligence? In the field of psychology, the debates volleyed between the definition of natural intelligence, referring to thinking or abstract reasoning, and the ways in which these can be measured (Legg & Hutter, 2007).

Attempts to measure intelligence introduced a range of intelligence tests in both humans and animals. When referring to human intelligence, the concept of IQ is one of the ways in which general intelligence factor, or the g-factor, is measured (Goertzel & Pennachin, 2007).

In their research, Legg and Hutter collected a variety of informal definitions of human intelligence provided by experts, distilled to the essential features which were then mathematically generalized as a measure to apply to machines (2007). Their informal definition espouses three key elements: an agent, environment and goals. The interaction

between each of the elements are referred to as actions and perceptions, depending on the goal. And since goals are generally challenging to express and are limited to just having one goal at a time.

To solve this, Legg and Hutter combine this expression and call it a reward so that any goal is simplified into an agent's way of maximizing the reward it receives. They also consider the environment, the space, intelligent agents and other variables in the equation. However eloquent this definition is expressed in precise mathematical terms, the application of such a standard definition as put forth by Legg and Hutter may not be easily operationalized in the context of policy-making around AI. Because meaning is inherently context-dependent, as Dreyfus argues, it cannot be formalized into discrete representations for computational programming (Susser, 2013).

Wirtz et al (2019) build upon Legg & Hutter's work by deriving a definition from extant literature. Analyzing six definitions, the authors remarked that the definitions represent both machine-based and human-like intelligent behavior. In particular, a key feature is AI's ability to imitate human cognitive behavior such as learning and problem-solving. AI therefore, according to them, refers to the "capability of a computer system to show human-like intelligent behavior characterized by certain core competencies, including perception, understanding, action, and learning" (Wirtz et al., 2019, p. 599).

Despite the authors' integrative definition of artificial intelligence, the ambiguity surrounding what artificial intelligence is continues to challenge researchers, practitioners, policy-makers alike as there is still no universally accepted definition available for it (Grosz et al., 2016). Consequently, the lack of a suitable foundation for defining AI, has implications on policy-making (Krafft et al., 2019) in addition to the social and ethical issues related to the autonomous decision-making capabilities of AI (Perri 6, 2001).

Researchers have grouped artificial intelligence into different categories: artificial narrow intelligence, artificial general intelligence and artificial super intelligence (Adams et al., 2012; Thierer et al., 2017; Wirtz et al., 2020). Those in the artificial narrow intelligence category are focused on one cognitive function, problem, or ability at a time, for example, a game of chess or voice assistants.

One real-life example is IBM's Deep Blue system that beat world champions in a game of chess (Adams et al., 2012). In essence, such applications can do one thing very well. These could include logical-inference based AI, simple algorithms, robotics, or expert systems (Goertzel and Pennachin, 2007). They may have minimal ability to communicate with humans and may sometimes be able to deal with unforeseen situations or discern

data patterns. They may sometimes require human expertise in their development or demand increased computing power to function (Goertzel and Pennachin, 2007).

Artificial general intelligence, on the other hand, has the ability to generalize its capability and learn on its own. Such systems will be capable of extending its intelligence and applying it in other fields. They will also have the capability to apply these characteristics onto other tasks without requiring human interference, in contrast to narrow AI systems, which lack these characteristics (Adams et al., 2012). These systems can be deemed closer to "strong" AI due to their ability to generalize, which is also a characteristic of human intelligence (Fjelland, 2020).

Research interest in this area has been growing thanks to recent innovations involving big data. Big data can refer to the problems or the techniques used to collect and maintain massive amounts of data for analysis. Big data can also refer to the use of mathematical models on large data sets in order to find patterns (Najafabadi et al., 2015). Simply put, big data are datasets that cannot be analyzed by humans without the aid of a computer (Thierer et al., 2017). When placed in the context of AI, researchers Zhuang, Wu, Chen & Pan suggested that integrating big data into machine learning complemented with human knowledge such as intuition can lead to "explainable, robust and general AI" (2017, p. 3) . Their work encourages prospective research directions in the creative ability of next generation AI, which they refer to as AI 2.0 (Zhuang et al., 2017).

Alongside the reinvigorated research enthusiasm in this area of AI, Dreyfus' early arguments concerning the impossibility of achieving artificial general intelligence are also rekindled. Fjelland maintains Dreyfus's argument that human knowledge is partly tacit, and thus cannot be expressed into a computer program (2020). Fjelland argues that although big data and deep learning present new approaches to achieving general AI, they will not be able to realize general AI in principle (2020). These breakthroughs, which were similar to the advent of neural networks, cannot realize artificial general intelligence in the next decades. A critical voice in the field, Fjelland strongly concludes that artificial general intelligence is a project that is" a dead end" (2020, p. 3).

Beyond artificial general intelligence, and perhaps even emerging from this, is artificial super intelligence. Artificial super intelligent systems will be more advanced, even expected to supersede the human mind (Wirtz et al., 2017). These are systems that are intellectually superior to humans and are thus classed as "strong AI." However, its development is not guaranteed to happen for several decades (Thierer et al., 2017). Etscheid goes on to suggest that strong AI systems are in the realm of science fiction (2019).

Despite developments in the field, a majority of the AI systems implemented and in practice will most likely be classified as “weak” given their specially adapted function suited for a particular purpose (Thierer et al., 2017, p. 9).

A number of international organizations have offered definitions to address the ambiguity regarding the lack of a standard definition for what is meant by artificial intelligence when developing policy in the field. Such organizations include the European Commission, the Organisation for Economic Co-operation and Development (OECD), the United Nations Educational, Scientific and Cultural Organization (UNESCO), and the Institute of Electrical and Electronics Engineers (IEEE). Some of these organizations are researching governance of AI and contributing to the development of policy in the area of AI (Wirtz et al., 2019). In particular, the European Commission, as of April 2021, presented a proposal for regulating AI. Looking at this definition, Article 3 of the proposal defines AI in the following terms:

> “artificial intelligence system’ (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with;” (EU Proposal AI Regulation, 2021, p. 39)

The techniques and approaches the article refers to are (EU Proposal AI Regulation, 2021, p. 39):

a. Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;

b. Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;

c. Statistical approaches, Bayesian estimation, search and optimization methods.

The European Commission states that by clearly defining AI, it can maintain the ability of ensuring legal certainty. At the same time, appending the annex provides some flexibility in order to accommodate future changes that may impact the regulation (Law - Point 6 preamble).

The European Commission, in this regard, addresses the ambiguity of the standard AI definition by future-proofing its own. Other organizations’ definitions are patterned in a similar vein - see Table 1. The OECD’s definition, for example, includes words such as

"human-defined objectives," "predictions," and "environments." Indeed, the OECD definition was drafted with the intent to recommend legal instruments for regulating AI.

Table 1. AI definitions by international organizations

Organization	Definition	Source
European Commission (EU Proposal AI Regulation, 2021, p. 39)	'artificial intelligence system' (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with; Annex I, Page I a. Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning; b. Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems; c. Statistical approaches, Bayesian estimation, search and optimization methods	AI Regulation Annex I, Page 1
OECD (2019)	AI system: An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments.	Recommendation of the Council on OECD Legal Instruments

	AI systems are designed to operate with varying levels of autonomy.	Artificial Intelligence
UNESCO (2021)	AI systems as technological systems which have the capacity to process information in a way that resembles intelligent behaviour, and typically includes aspects of reasoning, learning, perception, prediction, planning or control	Draft text of the Recommendation on the Ethics of Artificial Intelligence
IEEE (2019)	Artificial intelligence involves computational technologies that are inspired by – but typically operate differently from – the way people and other biological organisms sense, learn, reason, and take action.	IEEE Position Statement - Approved by IEEE Board of Directors 2019

Table 1. Definitions of AI

Unlike earlier suggested definitions for AI, the definitions in Table 1 expands beyond human-like cognition to more of a description of the different types of AI technologies. These definitions attempt to describe what AI can do rather than define the nature of artificial intelligence itself.

While these descriptions may not necessarily put to rest the debate over a universally accepted definition of what AI is, it does, however, provide a baseline for establishing general definitions that can then be used by policy-makers in AI regulatory context.

Because this research paper inquires into the state of AI ethics in practice within European context, it adopts the definition established by the European Commission in its proposal for an AI regulation. Hereto, AI can be any "software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with" (European Commission, 2021, p.39).

2.3 AI in the public sector

Artificial intelligence brings a wealth of changes that will impact society in the years to come. On one hand, the private sector is deploying expanded use cases for AI beyond customer service in various industries such as healthcare, retail, automotive, financial services, education, and travel (Sun & Medaglia, 2017). AI systems are deployed in ways that help companies target ads, detect fraud through predictive analytics, guide customers in navigating websites, and in some cases provide insights as decision-support tools for workers (Desouza et al., 2020; Mehr, 2017).

On the other hand, the public sector can potentially benefit from these applications as well but generally lags the private sector in AI deployment (Mehr, 2017; Berryhill et al., 2018). For the public sector, AI is said to have the potential to enhance the quality and consistency in delivering public services, improve policy design and implementation, reduce costs, increase security and facilitate interaction with citizens (Abbas et al., 2019; Chen et al., 2020; Desouza et al., 2020; Misuraca et al., 2020).

For example, the Wuhou Administrative Approval Bureau in China introduced an AI-based self-service machine featuring both face and fingerprint recognition and natural language processing for citizens to use (Chen et al., 2020). Fully equipped with an ID card reader, live camera, and QR code scanner, the machine offers citizens the ability to print government-issued documents such as social security certificates, licenses, and permits. The AI-based machine has provided support to about 30,000 users (Chen et al., 2020). In this context, this use case illustrates how AI can facilitate interaction with citizens.

The primary application of AI in public services is the reduction of the administrative burden (Mehr, 2017). This not only reduces cost, but in turn, frees up time and resources, allowing public servants to focus on more important, specialized work (Berryhill et al., 2019). Because public servants often perform repetitive tasks to process documents and conform to administrative procedures, AI can take on these tasks. Mehr categorizes the ways in which AI applications reduce the administrative burden into five procedures: answering questions, filling out and searching documents, routing requests, translation, and drafting document" (Mehr, 2017, p.6). For example, local prefectures in Japan have piloted chatbots to deliver public services in different domains. Normally a duty assumed by a public servant, the chatbots are deployed to respond to citizens' inquiries in the areas of waste collection and treatment, tax consultation, parental support and general information desk (Aoki, 2020). Given the nature of administrative procedures, by applying narrow artificial intelligence, which are well suited to perform specific, redundant tasks, there is potential to realize efficiency in this regard (Etscheid, 2019). In

this case, chatbots, a form of narrow AI, are used to answer general inquiries from citizens (Aoki, 2020). Routing requests have been experimented by the Mexican government where AI is used to categorize citizens' petitions and forward them to the appropriate office that will handle these requests (Mehr, 2017). Reducing the administrative burden on public servants is a way to achieve efficiency in delivering public services.

Core to the decision to deploy AI in the private sector is the financial incentive for performance efficiency, which creates conflicting stances with the ethical demands of its customers (Slee, 2020). Whereas for the public sector, the main driver for adopting AI solutions is increased effectiveness and efficiency. A systematic review performed in 2015 coalesced empirical research on public sector innovation. In their investigation of 181 articles, Vries, Bekkers and Tummers analyze definitions of innovation, its types, goals, antecedents, and outcomes. Their research shows that one of the main goals of public sector innovation is to increase effectiveness, followed by efficiency (Vries et al., 2016).

Extending this to AI-enabled innovation, AI is a form of public sector innovation when applied in this context. Therefore, AI can be expected to increase effectiveness and efficiency. This assertion is supported by a study in 2020 by Misuraca et al who undertook the mapping of AI use cases across the European Union. For 75 out of the 85 AI-initiatives surveyed, they noted that most AI projects in government are initiated with the objective of achieving efficiency goals. A number of these initiatives focus on generating internal efficiencies in order to improve organizational performance (2020).

As trends in big data and digitalization continue to grow, public service organizations are devoting resources to harness the power of data held within their organizations (Misuraca et al., 2020). Underpinning this drive for AI-enabled innovation is data governance. Based on their research, high-quality data is regarded as an antecedent for AI-enabled innovation (van Noordt & Misuraca, 2020). Data-sharing within public service organizations, while ensuring security and privacy, encourages AI development.

However, obtaining high-quality requires time and considerable resources. It is not uncommon to find data scattered throughout organizations, and as a result, the responsibility for its oversight and management is often ambiguous (Janssen et al., 2020). To this extent, research in this area points to the critical role of data governance in AI development.

Lastly, a systematic analysis of existing literature shows that AI collaboration between the private and public sector has shown both advantages and disadvantages (Reis et al., 2019). On one hand, partnerships with private companies enhance the delivery of services

to the public. On the other, it is unclear just how AI is transforming digital governments. Governments in Europe have led the way in the collaborative partnerships between private sector technology companies and government agencies. However, this approach have also shown some cons in regards to management that may influence the ourcomes of AI initiatives (Reis et al, 2019). Their study addresses the synergy between public and private sector that encourage AI uptake.

2.4 Challenges and risks

Despite the recent developments in the field, adoption of AI slower in the public sector (Zuiderwijk et al., 2021). Researchers have noted that there has been limited empirical research on artificial intelligence in the public sector (Sun & Medaglia, 2019). As a result, little is understood about the specific challenges of AI in the public sector (Aoki, 2020; Siau & Wang, 2020; Wirtz et al., 2020).

Inherent differences exist between public and private sectors, namely in value drivers, risk appetite, stakeholders, and goals. In their research, Desouza, Dawson and Chenok suggest that for the public sector, these challenges can exist in the following ways (2020):

- That the public sector must deal with complex political, societal, legal and economic factors that may not be applicable to the private sector.
- That AI projects for the public sector must serve the public good and deliver value
- That such projects must go beyond being cost efficient and cater to the diverse stakeholders and their conflicting agendas
- That decision making and operations need to be transparent and fair; and
- That since public sector AI projects are financed from taxes, these projects may be subject to oversight and inspections

These differences provide indications to the challenges that exist specifically in the public sector and have been flagged by other researchers as well (Zuiderwijk et al., 2021). A number of researchers have elucidated the challenges that public service organizations face with regards to AI.

Considering the scholarly debates that ensued on the topic, Wirtz et al propose four dimensions of AI challenges, namely: AI technology implementation, AI law and regulation, AI ethics, and AI society (2020) as shown in Figure 1.

Under technology and implementation, Wirtz et al describe multiple challenges facing the public sector. One of which is AI safety in the context of not just security issues, but the AI's having the ability to learn, and perhaps learn negative behaviour due to its surroundings. Following AI safety are concerns for data quality and systems integration. They describe this challenge in terms of data management, which leads to possible bias or inaccurate outputs. In addition, financial feasibility plays a role as budgets for innovation and developing AI solutions may cost the organization significantly as well as affording AI expertise which the public sector may lack.

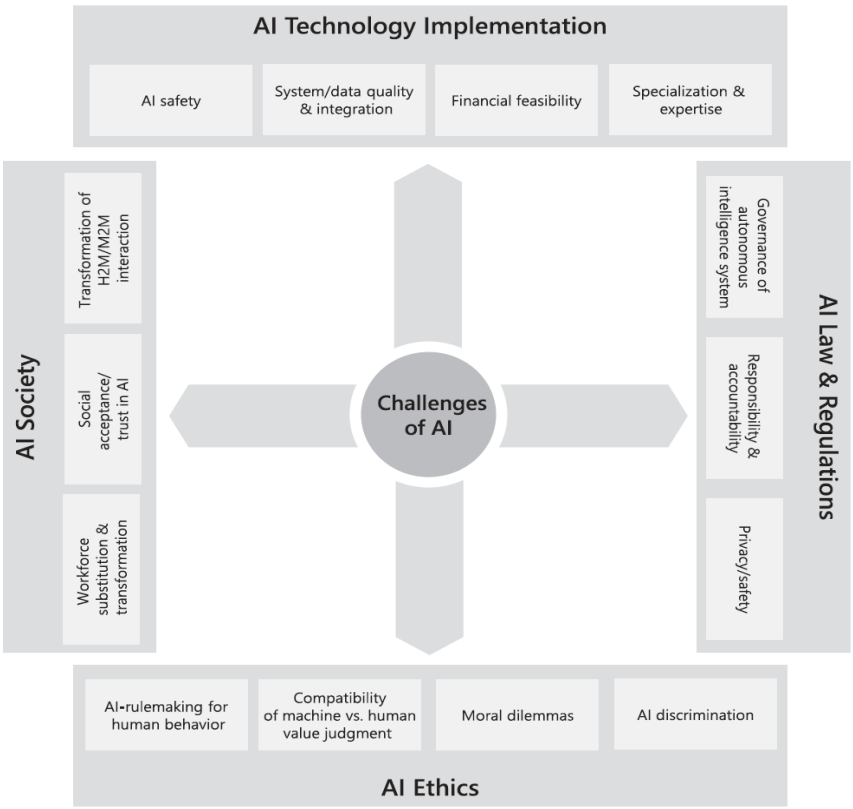

Figure 1. Public sector challenges of AI, based on Wirtz et al., 2020

The legal aspect of AI poses several challenges to the public sector, in particular questions on responsibilty and accountability. When failures occur, who is responsible for the decisions made by the AI? Answers to these questions and many others such as discrimination moral dilemmas become an important consideration for policymakers and is an often contentious area of discussion in the ethics of AI.

For society, AI threatens the labor markeets in terms of substituting economic opportunities for millions of people. Wirtz et al state that the implementation of AI could lead to unemployment due to automation. Governments and policymakers, as a result, need to consider how to address these impacts on the economy and societies worldwide (Wirtz et al., 2020).

Wirtz and Muller have expounded the risks of AI to the public. In their research, they highlight four prospective risks that public sector organizations should pay heed (2019). The first is the loss of control over systems. When large AI systems are interconnected, there is the risk that oversight of the millions of processes and transactions becomes

convoluted to a point beyond human control. The second risk mentioned relates to the authority granted to the AI without human involvement. They describe the worst-case scenario of an AI acting beyond the limits and making decision without authorization. Third, autonomy becomes lost as daily decision-making, included those of ethical import, is supplanted by the AI. And fourth, they describe the overreach of public safety and enforcement agencies when dealing with surveillance and breach of privacy (2019). While their research contributes to the broader understanding of the risks posed by AI, their methodology is not based on empirical studies, but rather a stream of literature on the subject.

2.5 AI ethics

Advances in AI and robotics have stimulated awareness and interest on the risks and challenges of AI. Because these risks are embedded in all levels of AI development - from the design of the AI application itself to its implementation for citizen's use, the ethics of AI becomes an important factor in terms of what the society would look like in the future (Bartneck et al., 2021). In particular, when decisions are made by autonomous systems, how should the AI system make such decisions? In other words, what decisions are the right and acceptable for a given set of variables under certain scenarios? As a result, a key issue in the field is defining to which ethical standards AI should adhere (Daly et al., 2019).

Before diving into the ethics of AI, it is important to lay the foundation for what is meant by ethics. In philosophy, ethics is a broad term encompassing multiple kinds such as descriptive, normative, deontological, virtue, applied, and so forth. Often, it is used interchangeably with morality (Bartneck et al., 2021). To some, ethics has been considered as a "soft law," that which is not strictly required by law but organizations follow to influence their brand image or manage their reputation in the media (Bartneck et al., 2021).

In existing literature, the ethics of AI concerns the moral obligations and duties of the AI and its creators (Siau & Wang, 2020). Siau and Wang suggested that understanding the ethics of AI can lead to the building of ethical AI. Therefore, it is crucial to have these discussions now and emboldens different stakeholders to carefully consider the ethics and associated morality of AI. Both features of AI, such as self-learning and autonomous decision-making, and the human factors involving accountability, standards, human rights, raise the societal, ethical issues associated with AI. They, therefore, recommend placing AI ethics at the center when developing AI and not as an "afterthought" (Siau and Wang, 2020, p. 84)

Some scholars, however, argue that AI systems are just another artifact no different from those before it such as factories and advertising that can instead help humans rationalize ethical decisions (Bryson & Kime, 1998). Although dating back two decades ago, their arguments hold substance to current questions facing AI and ethics. To them, such artifacts indicate society's uncertain ethics and the over-identification with machine intelligence. They assert that the nature of ethical obligations itself is often misconstrued. It is not the artifact -the AI, the calculator, or the pulley - but rather the similar flaws, errors and fallibilities that humans can make that against which precautions should be taken. They state that the real danger of these systems is through their misuse by the people who control them. They propose that AI programs can be used as a tool to help

humans understand the nature of relations within cultures and society that could be ethically significant. This can help society keep pace with the cultural shifts happening (1998).

Since there is still a lack of agreement on what decisions autonomous AI systems should make (Bartneck et al., 2021), a number of studies seek to understand how different societies determine what is ethical for AI. For example, the Moral Machine is an experiment deployed via an online platform that explores the moral dilemma involving autonomous vehicles (Awad et al., 2018). It collected 40 million decision input from millions of people in 233 countries. Their findings suggest that a universal agreement may not be attainable because even when strongest preferences were expressed, substantial cultural variations persisted. And while attempts have been made at creating ethical codes for AI, those that advocate for those aligned with human values, these codes do not fully account for inner conflicts, disagreements and cultural dissimilarities in morality of humans. Nevertheless, this experiment suggests that even with these factors in play, they may not necessarily result in fatal outcomes (2018).

As a result of the increase in uptake of AI, the debate around the risks and implications of AI have ushered a wave of societal concerns. Various organizations at different levels - national, international, supranational - have formed expert groups to work on these issues (Jobin et al., 2019). One expert committee, for example, is the High-Level Expert Group on Artificial Intelligence, which comprises experts and leaders both in the private sector and public sector. This committee was institutionally appointed by the European Commission to produce guidance and reports on AI. Because this area is burgeoning, expedient attempts to address these concerns have manifested through a number of different ways such as ethical guidelines, framework, standards, as well as regulatory proposals for AI.

Ethical frameworks and guidelines have cropped up around the globe to hedge the risks and implications of AI. In a mapping study of the global landscape on the guidelines for AI, researchers note that the majority of ethics guidelines are released in the United States and the European Union, followed by the UK (Jobin et al., 2019). Because Jobin et al's study of ethical guidelines around the world seeks to understand the convergence and divergence of what ethical AI should be.

Their study showed that of the 84 documents on ethical AI guidelines, a convergence appeared on ethical principles of transparency, justice, non-maleficence, responsibility and privacy.

A bibliometric study of AI ethics literature identifies four issues related to AI ethics: AI techniques, the implications in the political and technology arena, data privacy and more specifically in healthcare (Zhang et al., 2021). For example, the authors discovered that machine learning as an AI technique raises most concerns about ethical issues of fairness, liability and fraud. This concern is further extended into technological and political implications that relate to sustainability, responsibility, and digitalization. Attention to privacy has also increased as a result of the big data trend, for example social media, and thus raised awareness among the public.

However, critically important is the divergence that is observed in AI ethics, namely on how ethical principles are understood, how they are important, what issue or actors they apply to and how they should be put into practice (Jobin et al., 2019). These concerns point to the lack of clarity on which principles should be prioritized and how these conflicts be resolved. Thus, the discussions continue on how best to address these concerns in practice.

2.6 AI ethics in practice

Wirtz and Muller suggest codifying ethical AI standards and regulations and to monitor their enforcement to reduce the risks described in literature (2019). Doing so addresses the risk of losing autonomy in decision-making and preserving the freedom of choice. This suggestion is echoed in other areas of literature. For example, Jobin et al suggest that alignment of ethical principles at the technology governance level can be achieved through standardization of them (2019).

Standards being developed by the Institute of Electrical and Electronics Engineer on the Ethically Aligned Designed Initiative are an example of this. Yet, Jobin et al raise the question as to whether these policy instruments have impact on the practical implementation of AI or by the stakeholders upholding them. Particularly, do AI developers apply AI ethical guidelines in their practice?

Researchers note that abidance with principles outlined in ethical guidelines is poor in practice (Hagendorff, 2020). Research done by McNamara et al in 2018 found that instructing software engineers to consider a code of ethics do not have a considerable, observed effect in their ethical decision making. Thus, the onus of ethical decision making does not solely rely on the individuals. The software developers, programmers and AI practitioners, a study found, may not be rigorously trained on the ethical issues of AI nor are they supported by the organization to speak about ethical concerns (Hagendorff, 2020). In this regard, researchers call for exploration of the role of organizations in applying ethics in AI development (Ryan et al, 2021).

Wirtz and Muller recommend setting up a public AI ethics committee to monitor the practical implementation of these standards (2019). Jobin et al echo this suggestion but through ethical review boards. They assert that ethical reviews boards will play an important role in determining the integrity of ethics upheld in AI applications. They note however that while this could be useful, unless given proper authority, these independent review boards will most likely have difficulties in overseeing AI applications by private and public institutions (2019).

Practical implementation of ethics has been met with significant challenges. Empirical studies have shown that ethical issues are related to the immediacy on the temporal horizon (Ryan et al., 2021). This study explains that for the short-term issues, those that can easily be addressed by putting technical mechanisms in place have been acted upon. These issues can relate to security and privacy. Whereas longer term issues deal with deeper ethical questions on justice and fairness.

Empirical research by Vakkuri et al. on the state of ethical AI in the private sector practice, in which they surveyed over 211 software companies, shows that ethical guidelines do not seem to have observable impact in practice (2020). Furthermore, a question on liability indicates that about 40% of organizations have little familiarity with handling liability concerns of AI. On the topic of accountability, researchers doubt how effective decision-making processes are documented and tracked by the companies based on the descriptive responses to the surveys. Lastly, interpretative differences appear on transparency in the sense of system development. Even more telling is the view on being transparent to public authorities: about 19% of the companies responded that they do not know. For the private sector, companies may be able to shirk this responsibility on transparency. However, for public service organizations, transparency becomes an inherent obligation and moral duty (Jørgensen & Bozeman, 2007)

AI ethics should be seen as a process and not a technological solution. Technical fixes abound in ethical guidelines but do not propose technical definitions or explanations (Hagendorff, 2019). Indeed, Mittelstadt questions the logic that inadequate consideration of ethics thus leads to poor design and negatively impacts users. This author states the risk involved with oversimplifying concepts to be technically feasible to implement but ring hollow of the ethics involved. The study explains that AI ethics lacks a reinforcement mechanism and that deviations do not have consequences, which is supported by Mittelstadt's findings.

Hagendorff's view emphasizes the lack of enforceable mechanism in self-governance codes, and that this leads to developers being less willing to build AI other than improving their public image or trust. Hagendorff argues that these motivating factors often serve the economic advantages and are used in the private sector as marketing strategy (2020).

On a macro-level, regulatory action as a stronger form of governance for AI have begun to appear as nations conceive their national artificial intelligence strategies. Smuha's article examines the "regulatory tool box" that regulators can use when considering forms of AI regulation. However, the author states that one challenge regulators face is that they are also subject to the self-governance elicited by ethical frameworks minus the lack of enforcement (Smuha, 2021).

The European General Data Protection Regulation (GDPR) offers some coverage relevant to AI, particularly in processing personal data. The GDPR is considered as one of the most successful examples of values convergence. The GDPR is enforced not only in Europe but has scope beyond European borders, wherever European citizen's personal

data is processed (General Data Protection Regulation, 2016). In this aspect, the GDPR provides a template for how convergence can be attained when applied to AI ethical principles.

Indeed, the European Parliament and Council have paved the way in terms of the first AI regulation. As of April 2021, has released a proposal on AI regulation (EU Proposal AI Regulation, 2021). The measures in this proposal show consistency with existing policy provisions, particularly with the GDPR. It also aims to harmonize the rules on AI in order to improve the AI ecosystem, and in general the economic markets. The proposal includes a methodology for determining what AI applications are high-risk and that “high-quality data, documentation, traceability, transparency, human oversight, accuracy and robustness” are strictly required ((EU Proposal AI Regulation, 2021, p.7).

2.7 Case Context: AI in Estonia

Estonia has been selected as the country for this multiple case study because of its unique position as a highly digitalized society and a world leader in digital public services compared to other countries (European Commission, 2019, p. 12).

In the European Commission's Digital Economy and Society Index of 2019, Estonia consistently ranked among the top in digital public services for three years (2019, p.12). These accolades are made possible by the robustness of their digital infrastructure, the backbone of which is known as X-Road.

X-road is the data layer exchange developed in 2001. Data exchanged through X-Road is secured. Encryption of outgoing data and authentication of data incoming are the norm on this data exchange bus (Robles et al., 2019, p. 72). Over 900 organizations are connected daily through X-Road (European Commission, 2019, p.12).

In July 2019, Estonia introduced its national artificial intelligence strategy. Their AI strategy is a collection of actions that the Estonian government has undertaken to increase the uptake of AI in both the private and public sectors and within the government at all levels (Government of the Republic of Estonia, 2019, p. 1). Top three on the list of measures for the public sector alone is to introduce ideas and existing AI solutions, facilitate the development of AI projects by agencies, and financing research on the implementation of automatic AI-based decision-making support. Following this list are actions on data governance, consent management, principles for responsible use of data and increasing the availability of open data (2019).

The strategy also takes into account the legal environment for AI to flourish. Measure number 4.1 states: "There is no need for fundamental changes to the basics of the legal system, but there are some changes in different laws to be made" (2019, p.10). This statement hints at the broader attitude towards regulation of innovation. For them, legislation could have the potential to hinder innovation and should cover the nature of transactions and sensitivity of data instead.

Estonia has legislation in place related to governing its digital society, therefore, legislative efforts should concentrate on analyzing existing laws and identify gaps. A study was done in 2020 on the legal dilemmas of AI in Estonia. In this study, the legal expert group responsible for AI regulation in the nation found that a separate AI law is not possible or needed, citing that AI is and remains a tool for humans and that it carries the will of the human being (Kerikmae & Parn-Lee, 2020).

This brief background on Estonia's AI strategy serves to inform about some of the priority and conditions that enabled the AI ecosystem to thrive and see the rise in AI use cases across its public administration.

3 Theoretical Framework

3.1 Value Sensitive Design

Based on a recent analysis of available literature, theoretical development for AI in the public sector is limited. As such, there have been calls for additional research incorporating interdisciplinary perspectives for AI (Zuiderwijk et al., 2021). Because of this, the Value Sensitive Design (VSD) serves as the theoretical as well as methodological framework for this research. In the past decade, there have been only a handful of studies that have espoused VSD in AI research (Umbrello & De Bellis, 2018; Umbrello & van de Poel, 2021; van Wynsberghe, 2013).

VSD is particularly applicable to this research because the approach integrates values into technical system design. It has been used in the context of technology and more advanced technologies such as AI, in particular robotics in healthcare (van Wynsberghe, 2013).

VSD draws from the human computer interaction field and embraces the sociotechnical approach. Perspectives of sociotechnical systems view the mutual shaping of society and technology, and how human values and technology enmesh (Cummings, 2006). Human values are an integral part of discussions surrounding the ethics of AI. The human values with ethical concerns that are often described in AI literature are fairness, justice, autonomy, privacy, trust, accountability, among others (Floridi et al., 2018). VSD, therefore, is an extensible and adaptable framework to support the inquiry into the state of ethical AI in practice because it addresses the human, ethical values embroiled in the ethical AI discussion.

VSD is a term coined by Friedman, Khan and Borning (2002). It is a "theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" (Friedman et al., 2008, p. 70). They refer to "value" as what a person or group of people consider important in life" (2008, p.70). Values, Friedman et al acknowledge, are not molded only by experiences of the world, but that they also depend on human interests and desires within a cultural context, an assertion in accordance with philosophical and sociological arguments by Behar, Bryson & Kilme, and Perri (Behar, 1993; Bryson & Kime, 1998; Perri 6, 2001).

In addition to being used as a theoretical framework or system design approach, VSD has a characteristically tripartite methodology that combines conceptual, empirical, and technical investigations as shown in Figure 2. Value sensitive design tripartite methodology (Friedman et al., 2008). Since VSD is also iterative, the investigations need not happen independently, and may overlap.

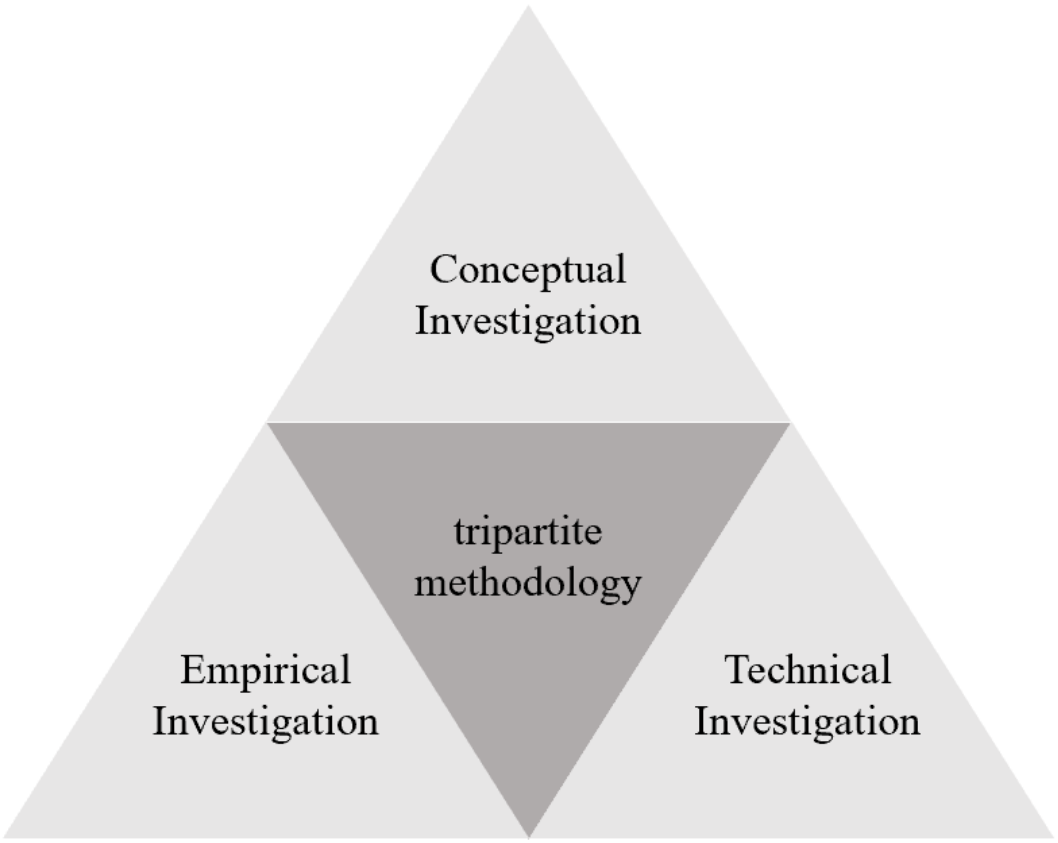

Figure 2. Value sensitive design tripartite methodology

The conceptual investigation is two-fold. On one hand, it explores the value source, implications, trade-offs in a technology's design. Questions such as "what values should be supported in the design process?" or "how are values supported or diminished by a particular technological design?" are some points to pose during this investigation (Friedman et al., 2002. p.2) On the other hand, conceptual investigation involves the thoughtful, sometimes philosophical consideration of all the direct stakeholders involved as well as indirect stakeholders that may be implicated by the values and technology. It is for this reason that VSD is a useful approach in determining the potential impact of an AI system and the human values implicated in its design and use.

The empirical investigation concerns the examination of the stakeholder's "understandings, context and experiences" relative to the technology and values (Friedman et al., 2002, p.3). Some suggested questions are "how do stakeholders apprehend individual values in the interactive context?" and "how do they prioritize competing values in design trade-offs?" (2002, p.3) More importantly, this investigation attempts to ascertain that the values explicated by the stakeholders in the conceptual investigation are espoused through practice and in the technology.

Furthermore, the empirical investigation can also inquire beyond the designer and into the organizational context of the AI and stakeholders. For example, some questions to consider may be related to the organizational motivations, training methods, reward structures and incentives as well as how the consideration of values can lead to positive outcomes for the organization in terms of revenue, reputation, employee satisfaction and so on.

The technical investigation inspects the technological properties, mechanisms or features that may implicate identified values and stakeholders. It focuses on the technology itself. It can also encourage the proactive design of technology by embedding the values identified in the conceptual investigation (Friedman et al., 2002). Questions such as "how do parts or functions of the technology support or hamper human values?" or "What design trade-offs should be prioritized in the future to support explicated values?" (2002, p.4)

VSD Tripartite Methodology	Sample Questions
Conceptual Investigation	• What are values? • Whose values should be supported in the design process? • How are values supported or diminished by particular technological designs? • How should we engage in trade-offs among competing values in the design, implementation, and use of information systems (e.g., autonomy vs. security, or anonymity vs. trust)? • Should moral values (e.g., a right to privacy) have greater weight, or even trump, non-moral values (e.g., aesthetic preferences)?
Empirical Investigation	• How do stakeholders apprehend individual values in the interactive context? • How do they prioritize competing values in design trade-offs?

	• How do they prioritize individual values and usability considerations? • Are there differences between espoused practice (what people say) compared with actual practice (what people do)? • What are organizations' motivations, methods of training and dissemination, reward structures, and economic incentives? • How can designers bring values into consideration, and in the process generate increased revenue, employee satisfaction, customer loyalty, or other desirable outcomes for their companies?
Technical Investigation	• How do parts or functions of the technology support or hamper human values? • What design trade-offs should be prioritized in the future to support explicated values?

Table 2. VSD tripartite methodology sample questions to consider - adapted

VSD has a wide range of features that can be beneficial for conducting empirical research on AI ethics in the public sector. First, the tripartite methodology allows for the inquiry of existing values implicated in the design of an AI system as well as the proactive design of these values in future designs. In addition to this, the methodology is iterative and integrative; it can be applied early in the design phase and throughout the process (2008, p. 85). Second, VSD emphasizes the need to identify both direct and indirect stakeholders, who are often an afterthought, if thought of at all, in the design process (Friedman et al., 2008, p. 86).

In the context of AI, these two qualities are an important consideration when designing and developing AI for the delivery of public services. Third, it distinctly articulates explicated values and technology trade-offs, facilitating the identification and prioritization of these trade-offs by the stakeholders.

Lastly, Friedman et al suggest that because value, technology or context of use can be a core motivator through which VSD can be initiated, VSD claims that although certain values are universally held, there are some that differ relative to a particular cultural context and time period (2008, p. 86).

There have been a number of criticisms associated with VSD, in particular lack of concrete ethical commitment and claims of universal values (Davis & Nathan, 2015).

Davis and Nathan, for example, highlight in their paper that VSD draws various ethical theories, for example, deontological, consequentialist, virtue, to name a few, but does not commit to any one of them. However, Davis and Nathan also do not specify whether VSD must always be complemented with an ethical theory (2015, p. 33).

In regards to VSD's claim of universality of values, Borning and Muller rejects VSD's claims, calling it "enormously problematic", and its position on cultural relativism "problematic as well" (Borning & Muller, 2012, p. 1126). Instead, they suggest that VSD assumes a pluralistic and humble position that can then clarify "whether VSD is a method that can be applied to any set of values or that VSD is a methodological instantiation of a particular set of values" (2012, p. 1126). For Borning and Muller, the VSD can thus become more widely adopted when these claims are toned down.

3.2 AI4People Ethical Framework

Acknowledging the benefits and limitations of this approach, this research adapts the VSD approach by complementing it with the AI ethical principles or values that are raised in discussion of AI ethics. In their paper AI4People, Floridi et al synthesize five ethical principles that underpin the development and adoption of AI that serves the good of society (2018). Illustrated in Figure 2, these principles are beneficence, non-maleficence, autonomy, justice, and explicability.

These principles are sourced from various organizations and initiatives that have also outlined principles for AI. Some of these principles from organizations, among others for example, are derived from the Asilomar AI Principles by the Future of Life Institute in 2017, IEEE's General Principles described in version two of Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems as well as principles stated by the European Commission's Group on Ethics in Science and Technologies in 2018.

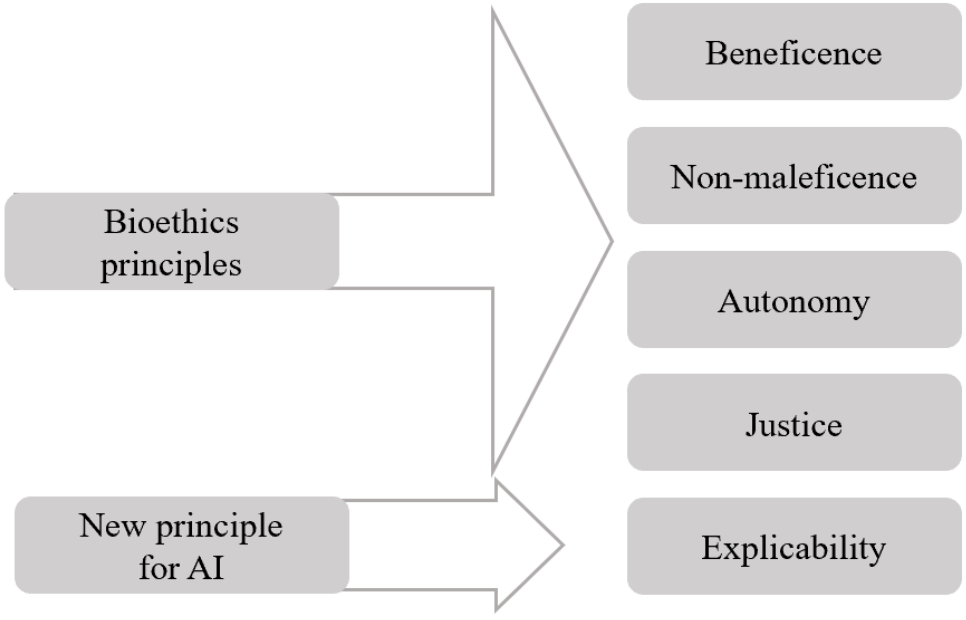

Figure 3. Ethical framework for AI, comprised of five principles (Floridi et al., 2018, p. 700)

Overall, there are 47 principles from which Floridi et al derived to produce the four core principles, which are used in bioethics: beneficence, non-maleficence, autonomy, justice, and explicability, an additional that they argue should be included. Floridi et al note that these principles minus explicability are similar to those in bioethics because this area of applied ethics deals with new forms of agents, patients and environments that AI ethics also faces (Floridi, 2013). Similarly, Jobin et al's research on the convergence and divergence of AI ethical principles in more than ethical guidelines studied suggests that there is a convergence around six principles, namely transparency, justice and fairness, non-maleficence, responsibility and privacy (2019, p. 391). These six are represented in AI4People's principles.

Beneficence: At its core, beneficence means promoting good in ethical terms (Jobin et al., 2019). Viewed as the common good, this principle concerns the promotion of well-being, preservation of human dignity, and sustaining the planet. Some strategies for progressing this principle is through alignment with human values and minimizing power concentration (2019). In this framework, the promoting the well-being of both humans and the planet are key to this principle.

Non-maleficence: According to their study, non-maleficence appeared more frequently than beneficence (Jobin et al., 2019). Privacy, security and safety are home to this principle. Privacy is closely related to management of personal data, including its access, use and control (Floridi et al., 2018). Security takes into account the mechanisms – often technical - in which privacy is preserved. In addition, the intentional and unintentional cause of harm falls under this question. Whether the harm originates from the AI itself or the humans involved in developing the technology remains unclear and thus contentious.

Autonomy: Floridi et al explain that in bioethics, autonomy refers to the idea that patients have the right to make decisions about receiving treatments that would impact them. In AI ethics, the parallel is seen when such decisions are delegated to AI agents outside oneself. Several ethical principles advocate for human's ability to choose and decide. Thus, this principle seeks to maintain the value of human choice (2018).

Justice: Under this principle are the concepts of equality, (non)-discrimination, accessibility, access and distribution, inclusion, fairness among others (Jobin et al., 2019). Floridi et al describe this principle as promoting prosperity and preserving solidarity. More precisely, their research indicates that justice can refer to a) using Ai to correct past wrongs, b) ensuring that the AI creates shared benefits, and c) preventing the introduction of new harms that exploit existing social structures (2018, p. 699).

Explicability: A number of values are expressed in this principle, in particular, accountability, transparency, comprehensibility and interpretability. This principle refers to the explicability of AI in the sense of being able to understand what it does and why it is making the decisions it makes and holding such decisions or processes to account. However, there are still significant discussions on whether AI systems should be held accountable the same way humans are or whether only humans are held accountable because it is they who are responsible for the AI system (Jobin et al., 2019).

Altogether, these principles work together to bring about an AI ethical framework. However, a critical voice emerges a year later. Mittelstadt states that although AI ethics has found convergence that replicate the four classic principles of medical ethics, he asserts that principles alone are insufficient to guarantee ethical AI (2019). He calls for regulatory action in order for the translation of principles into practice to be a cooperative process. Thus, he believes it is too early to celebrate consensus over these principles (Mittelstadt, 2019)

Heeding Middelstadt's criticism on principles, one of this research's goals is to identify the state of ethical AI in practice in the public sector, whether such principles are indeed being translated into practice, and if so, in what ways. While Floridi's principles may not fully represent AI-implicated human values, nor does it claim any universality, the AI4People principles are a solid foundation to aid in the inquiry of AI ethics in practice. Thus, for the purpose of this research, these two frameworks - VSD complemented with AI4People ethical AI principles - are carefully selected to facilitate in answering the research question despite some aforementioned limitations.

4 Methodology

This section explains the research methodology, which adapts the Value Sensitive Design (VSD) methodology. The following subsections explains the research design, collection of data and its analysis, as well as limitations of this research.

4.1 Research Design

To discern the how AI ethics is considered in practice by public service organizations, this qualitative study is guided by the VSD's characteristically tripartite methodology: conceptual, empirical and technical investigation. The tripartite methodology is apt for the purpose of uncovering the ethical values at play in the design of the AI use case, the experiences of stakeholders involved in its design, and the technical components of the AI itself. VSD focus on values of is extensible in that it also aims to predict values and issues that may arise throughout the design of technology (Umbrello and DeBellis, 2018). For this research, the tripartite methodology provides the pillars to support translation of ethics into practice:

The conceptual investigation's two components are applied in this research. First, participants and involvement of parties are inquired through the stakeholder analysis. This inquiry allows to draw the values that play a role in the design and development of the AI. Unlike other methods that ascribe roles and duties to a particular stakeholder (Umbrello & DeBellis, 2018), VSD's stakeholder analysis covers both direct stakeholders that were involved in the AI development as well as the indirect stakeholders that may be implicated by the design, development and use of the AI. Friedman et al state that indirect stakeholders are left ignored in the design process (2002, p. 3). Secondly, the identification of values is explored in this investigation. "What" values and "whose" values are important questions to consider in understanding the intent and motivations of the stakeholders in the design of the AI (Friedman et al., 2002, p.2). The nature of these questions seeks to identify the values that ultimately influence the AI.

The empirical investigation explores the extent of which individual values are apprehended in the context of AI design and development and the extent which these values are prioritized in deign trade-offs. This investigation elicits these values in the context of the AI, the stakeholders experiences, the issues and challenges that may have occurred and so on. In addition, this investigation takes into account the success and failures seen in development (Umbrello & DeBellis, 2018). Feedback from direct and

indirect stakeholders about the AI are captured under this investigation. The empirical investigation's unit of analysis is the people.

The technical investigation is straightforward and comprises the tangible properties and components of the technological artifact (Friedman et al., 2002). This investigation inquires into how these technical components support the identified values. Moreover, the technical investigation is forward-looking in that it can also discern technical components or mechanisms that preemptively support values in the conceptual investigation. The unit of analysis for this investigation is the technology alone.

However, because the set of values adopted by the VSD methodology have limitations (see Section 3.1), the prescribed values set is not well-suited to answer the questions of this research. As such, the VSD methodology is complemented with AI4People's five ethical AI principles: beneficence, non-maleficence, autonomy, justice, and explicability. These principles are more suitable for analyzing the ethical principles specific to AI. As such, it is this set of values that were explored in the values identification of the conceptual investigation.

This research focuses on Estonia as a country context of study due to its highly digitalized public services, its aggressive AI strategy and the extensive collection use cases of AI in the public sector. Estonia has over 70 identified use cases for AI in the public sector (Government of the Republic of Estonia, 2019). The AI use cases are designed and developed by public institutions ranging in function such as public safety, social welfare services, border patrol, health, transportation, finance, education, and so on. A large portion of these use cases are "in development" while a great number have already "been implemented." It is a suitable context to study for the purpose of understanding the state of ethical AI in practice.

Of the 70 use cases displayed on Estonia's AI strategy website, 8 have been selected based on the following factors:

- The AI use cases selected come from a diverse domain of public services.
- The AI use cases provide a service to the public, or aid in delivering a public service.
- The AI use cases interact with the public directly or the public is implicated by their use.
- The AI use case development status, whether in development or implemented, provide

In addition, the selected use cases were limited to the organizations that were available and agreed to this research on the condition of anonymity. The list of use cases is listed in Table 3 in alphabetical order.

No.	Public Service Domain	Use Case AI Type	Development Status
1	Administrative	Chatbot	In development
2	Administrative - IT	Chatbot, decision-support	Implemented
3	Education and culture	Facial recognition	Implemented
4	Finance	Risk scoring	Implemented
5	Public infrastructure	Forecasting and planning	Implemented
6	Public safety	Transcription and risk assessment	Implemented
7	Regulatory and oversight	Machine learning	In development
8	Social welfare services	Decision-support	In development

Table 3. AI use cases by public service domain

4.2 Data Collection

Qualitative data in the form of semi-structured interviews was collected from respondents from eight public service organizations that have developed an AI solution across the Estonian public administration. The respondents' roles varied organization to organization, however, the commonality was their direct involvement in the design and development of the AI solution. Their roles are indicated Table 4.

Open-ended questions structured according to the suggested VSD tri-partite methodology such as stakeholder involvement, design considerations, values implications and so on guided the interviews. The questions were crafted in such a way that captured the key inquiries of the VSD tripartite methodology while also enabling the investigation of AI4People's ethical AI principles. Respondents were given time to fully converse and answer in narrative form. Refer to Appendix B for the interview guide.

Respondent	Respondent's Role	Data Collection Date	Data Collection Format
R1	Data and AI specialist	05 March 2021	Semi-structured interview
R2	IT service developer	01 April 2021	Semi-structured interview
R3A	Development specialist	28 May 2021 07 June 2021	Written responses followed by a semi-structured interview
R3B	Technical procurement specialist	27 May 2021	Semi-structured interview
R4	Technology development specialist	26 May 2021	Semi-structured interview
R5	Data analyst	02 June 2021	Semi-structured interview
R6	Third party AI developer	27 May 2021	Preferred written-responses
R7	AI project lead	01 April 2021	Semi-structured interview
R8	AI product manager	26 May 2021	Email response

Table 4. Interview respondents' roles

In total, data was collected from 9 respondents representing the 8 public service organizations. The average length of interviews was 30 minutes conducted via video chat. The interviews were recorded and transcribed using an online transcription service. The transcriptions were reviewed for accuracy. Anonymity of respondents was respected. As such, a number of identifiable characteristics were omitted to preserve privacy and confidentiality.

4.3 Data Analysis

Coding was used to analyze the data collected. A code, Saldana explains, is a qualitative inquiry that captures the essential or salient points in language-based data (2015). Because the AI4People ethical AI principles are anchored in values, values-based coding was performed, and codes were categorized according to the VSD's tripartite methodology. The outcome of coding was grouped into themes that relate to AI4People's ethical AI principles. For example, codes related to firewalls, authentication, passwords were grouped under the category 'security,' which in turn was under the theme of 'technical investigation.' Another example is 'get help faster' which was categorized as 'efficiency' under the theme of 'conceptual investigation.'

This research involved multiple AI use cases. As such, each use case was coded individually before proceeding to the next. The electronic coding software MAXQDA was used to facilitate the coding process for multiple AI use cases. And because coding is cyclical (Saldana, 2015), the analytical process was rigorously iterated to ensure that themes emerged.

4.4 Methodological Limitations

The methodological approach of this research is subject to limitations that readers should bear in mind. First, due to the finite amount of time and resources, the scope of this research has been narrowed to a single country in the European Union and within that the public sector context in Estonia. Therefore, in terms of external validity, the applicability of the findings in this research may not be generalizable for other country context well beyond the borders of Europe which may be subjected to different measures, times, people.

Second, the unit of analysis is concentrated on the AI use case and the circumstances surrounding the design and development of the AI. Only one respondent was interviewed for each of the seven AI use cases, whereas two respondents from two different organization provided data for one of the AI use cases as this was a joint AI project (labeled R3A and R3B).

Consequently, the perspectives offered on each of the use cases are significantly limited to these respondents' perspectives and may not be reflective of the entirety of the AI project nor of the organizational whole. Furthermore, most of the AI use cases were not completely developed or in full operational use. Thus, the broader, more in-depth analysis could not be performed. However, the author strived to expand the number of case studies to provide robustness in this regard. For future iterations of this methodology, an in-depth, longitudinal or a single case study of a completed and deployed AI solution may yield more substantial insights to address on the research topic at hand.

Third, researchers have pointed to the limitations of VSD both from a theoretical and methodological point of view. These limitations have been explained in Section 3 of this research. However, in relation to this, the complemented use of AI4People's AI ethical principles may have constrained the range of ethical values that could have emerge from the analysis. Although the ethical principles do not purport universality, they have been systematically condensed to the five ethical principles presented originally fetched from reputable international and scientific institutions.

Fourth, indirect stakeholders were not included in the scope of this research, in particular, the citizens that may be implicated by the use of the AI. This component of the VSD framework was addressed by way of asking questions about feedback on the AI from the direct stakeholders. Therefore, their views and values were not represented in the conceptual investigation.

Lastly, the analyses of the transcriptions were performed by the author alone, and no additional analysts were involved in the coding of the transcriptions. The electronic coding software did not perform any analyses on behalf of the author; it was merely a tool used to assist in the organization and process of coding. Bias may have been introduced in the methodology, coding, analysis and therefore may affect the interpretation of results.

5 Results

Following the tripartite methodology, the interview questions sought to extract relevant information in order to sufficiently answer the main research question and two accompanying sub-questions, which are:

How do public service organizations ensure ethically-aligned AI public services in practice?

1. *What are the key issues that public service organizations face in AI design and development?*

2. *In what ways are AI ethical principles considered in practice by public service organizations in the design and development of AI for public service delivery?*

Because answering the research questions required obtaining an understanding of the context in which the AI use cases, humans and values are implicated in the design and development of the AI, the interview covered relevant aspects of the AI as a project. The results of the interviews are described in this section and are organized thematically under each of the three components of the VSD methodology - conceptual, empirical, and technical.

Researcher's note: Each AI solution is abbreviated with R for respondent, and a corresponding number indicating the public service organization such that Respondent from Organization 5 is shortened to R5. In addition, due to privacy and confidentiality, the names of individuals and organizations have been omitted.

5.1 Conceptual Investigation

Efficiency-related goals and objectives

In order to understand which values and whose values were at play when developing the AI system, questions related to the AI project's conception, background, objectives and the intention behind its development were posed to the respondents. When asked how the AI project came about, respondents shared the origins, drivers and motivations behind the use of AI. Using data to solve a problem was a common theme that emerged for most of the organizations, with the intent to improve internal processes or public services and make them more efficient.

R2 described the conception of the AI project as an "organic development." R2 shared that since their organization "always had good, well-structured data about clients, services and outcomes" that had been in place for more than 10 years, they decided to use their data in order to "target better the services" being provided to their clients.

The reasoning was repeated by R4 who observed that there were volumes of data that already existed within their organization. R4 researched ways in which data can be deduced from existing data in order to minimize the burden on citizens' providing this same data.

The availability of funding from R1, the head agency dedicated to supporting data-driven AI initiatives, cinched their commitment to piloting the AI solution. R4 remarked that they did not fixate on the AI as a solution solely for the main reason that there was funding available. Indeed, R4 emphasized that theirs was a problem that needed to be solved, and AI was a possibility for solving that problem.

Moreover, R4 highlighted the importance of understanding whether AI is needed for the problem at hand, the magnitude of the problem, and associated expenses. Next important consideration would be to determine the availability and quality of data by the organization.

Data-driven, AI workshops and initiatives at the head agency encouraged R3A's leadership to participate in a data analysis workshop. This participation meant compiling problem descriptions and discovering possible solutions to address the problem. Shortly after, the procurement process for the AI solution was underway, in which R3B's organization, among others, became involved.

For R5's AI project, the main driver was to assess the efficiency of measures being implemented and funding allocated for public safety. "We want to have such a tool where

we input some data, it looks at the data and the patterns, combinations in this data. And as an answer as an output it, it gets us these assessments." Similarly, R6 explained that the purpose of the AI solution is to gauge the extent of efficiency improvements within the organization's monitoring tasks to comply with law by leveraging AI/ML.

Interestingly, R7's AI project diverged from this common theme of efficiency and problem-solving using data. For them, their AI project was born out of one of their programmer's desire to develop skills and competencies in AI, specifically in image recognition. This experimental project later evolved into an AI facial recognition tool that offered "users some joy and entertainment and some practical value" (R7). The AI facial recognition tool allowed the individuals in the public to submit images, which were then matched by the AI with publicly available records, the results of which were returned automatically to the requestor via email.

Establishing maturity of AI solutions

The maturity of the AI solutions appeared as a consistent theme throughout the interviews because fo**r a majority of the organizations, the maturity of their AI solutions were at the early stages of development**. The AI solutions were described as a "proof of concept" (R3B, R5, R6), "a prototype" (R3A, R5), "trial phase" (R4), "a pilot" project or phase (R2, R4, R5).

R3A and R3B's joint AI solution was not in use even though development of the prototype was completed. R3B pointed to additional work needed regarding the technical specifications of the solution and some data-related concerns as reasons for this. R4, on the other hand, had completed their first phase of trials with the AI. R6 also built a proof-of-concept which was still in the development phase.

Generally speaking, only two organizations are using their AI prototypes for day-to-day use (R2 and R7). R2's AI solution was used internally to help management assess distribution of workload and determine risk scores for clients. R7's publicly-facing AI solution was deployed and available for use by the general public. It is very important to note that this AI solution did not facilitate the rendering of any kind of public service in the traditional sense, such as safety, welfare, education, health, and so on. This AI solution was developed out of "entertainment" but that which was developed for public use by a public service organization, hence its inclusion.

At one point, there was an organization (R8) that was also considered for the reporting of results. However, R8 explained that their AI solution, although publicly available for use,

is at a very early stage of development, and therefore some of the questions relevant to this research and ethical AI could not be answered. This paragraph was the only mention of this organization because of the ethical duty as a researcher to present this as a result nonetheless. Moving forward, this organization will no longer appear in the rest of this section because there were no responses provided to the interview questions.

Understanding feasibility

The early stages of development were critical for these organizations to ascertain the feasibility of developing the AI solution for solving the problem they had identified.

In-depth, discrete discussions were held project stakeholders to determine "what can be done and what cannot be done" (R3B). The development of a prototype helped them answer these questions. Featuring the most basic components required to function, the prototype allowed the organization to experiment while managing costs. R3B shared that "we don't want to make a high value, high cost solution with no effects."

With R4's piloting phase nearing its completion, they wanted to find out whether it was possible to use the AI. "And the answer to that is yes" (R4). Following the completion of the prototype, R4 had created reports and assessments, having established that the AI solution indeed worked. During the interview, it was shared that they were in the process of determining whether to continue onto the next phase of development.

For, R5 explained that their AI solution was "one of the first [AI] projects in our administration." And that for them, this is a proof-of-concept prototype "to answer the question, is it possible the thing that we want or not?" Understanding the extent to which the proposed AI solutions could solve problems or meet efficiency-related goals was a key activity for some organizations in the study.

Involving stakeholders

Identifying stakeholders, or those directly involved and indirectly implicated by the technology, is an element of the conceptual investigation phase. First, the respondents were asked questions in relation to the different stakeholders involved in the design of the AI solution. Secondly, the respondents were asked whether the potential impact of the AI solution on all stakeholders was considered.

As head agency for AI-related initiatives, R1's organization reviewed questionnaires submitted by organizations for any AI project that sought funding through them. R1 explained that their AI project framework requires organizations to identify different key stakeholders and the risks involved, particularly those affected by personal data and may fall under the purview of the GDPR and other regulations.

In addition, R1's organization provided guidance to "think through what is actually possible to do" when it comes to AI solutions. Afterwards, R1's organization would fund projects and assist in the tender processes. AI solutions developed by the organizations were either advised by (R3, R5, R6) or received funding (R4) from R1. The remaining projects in this research were not affiliated in any way (R2 and R7).

R2 listed the key stakeholders involved in creating their risk scoring AI solution, of which were the market researchers from the local university who worked together with internal teams in R2's analysis department on the logic behind the AI solution. Noting that R2 used their AI solution internally to manage workload and provide support on decision-making procedures to identify clients who require tailored assistance as part of a pilot phase, the end users, in this case, were the internal teams. Yet, the output of the AI solution directly impacted the clients whose input was not considered in the design of the solution. The rationale was that the AI solution was new and had too little data.

R3's AI solution was designed through a combined effort from various stakeholders from other organizations and a third-party vendor. However, the impact of the AI on the stakeholders, specifically the general public, was not considered because ultimately, according to R3A, a manager would make the final decision on the output of the AI solution, thus placing the onus on the human and not the AI solution. R3A shared that it was made clear that the manager or whoever was receiving information from the AI solution would be responsible for decisions made thereafter that would impact the well-being of a citizen or public being served. No specific regulation was referenced; however, this was agreed upon internally.

R3B agreed, stating that the AI solution "is not making decisions by itself." The same was said for R4, whose end-users were internal teams and the humans made the ultimate decision. R4 tested to make sure that the teams were able to arrive at a similar decision before approaching the client. "The only difference is we do this in excel or AI as a technical tool" (R4).

User feedback was solicited as was the case for R4 and R7. Feedback for R4 had been positive as the AI solution worked "very well" with their internal processes and teams, who were the users. Feedback from indirect stakeholders implicated by the AI solution

was not collected, such as the general public. Whereas user feedback was requested and collected by R7 through an online intake form, available along with the AI solution. However, feedback was not yet incorporated into the design of the AI solution when this interview was conducted (R7). It is worth noting that although not a direct stakeholder, R7's developer used an open-source AI solution to deliver the facial recognition service to the public.

One of the stakeholders heavily involved for R3, R5 and R6 was a third-party vendor of AI solutions whose mission was to solve private and public business problems. R6's third-party vendor declared that for them, there was "no conflict from an ethical point of view as the aim is to help the client enforce laws ultimately aimed at protecting [citizens]."

Transparency

Given that the AI solutions were used to help deliver a public service, thereby serving the citizens, a question about whether the citizens should be informed that an AI tool was involved in delivering the service. Only two of the AI solutions presented here are at the point where this question could be pertinent: R7 and R2.

R7's AI solution was open for the public to use on their "own free will." Thus, with this intention and consent, it would be obvious to users that they would be engaging with an AI solution to provide the service. Again, this AI solution did not facilitate or was meant for the rendering of any kind of public service. This AI solution was for entertainment purposes, which were communicated with the public.

R2, however, was meant for facilitating the delivery of a public service. Aside from the terms of data processing outlined in their policies, R2 the clients were not made aware that an AI solution was used by the organization to help deliver a service to them during the pilot phase.

Consultations with privacy specialists and legal counsels indicated that the organization was within legal grounds to process client data, Therefore, consent of the client was not required and thus clients were not informed. However, there were discussions about the best way to present the output of the AI solution to their clients without negatively impacting the client's well-being. R2 stated "we are thinking of how to do it, but this is something we are intending to do."

A number of the AI solutions were in early developmental stages, and as such, the question about transparency could only be answered in the hypothetical future should the

AI solution be fully deployed and used. Answers appeared to reach a consensus over whether the general public should be informed about the use of AI in delivering services to them. All of them agreed that the citizens should be informed about the AI's involvement in delivering a public service, regardless of whether it directly affected them. Of course, it was also discussed that the lay person may not have the skills to understand the complexity of AI. At the very least, the use of AI should be communicated to the public. Transparency, it seemed, tied directly to the inherent duty of public service. There was a strong sentiment that as public institutions serving citizens, they should be open and transparent about these matters.

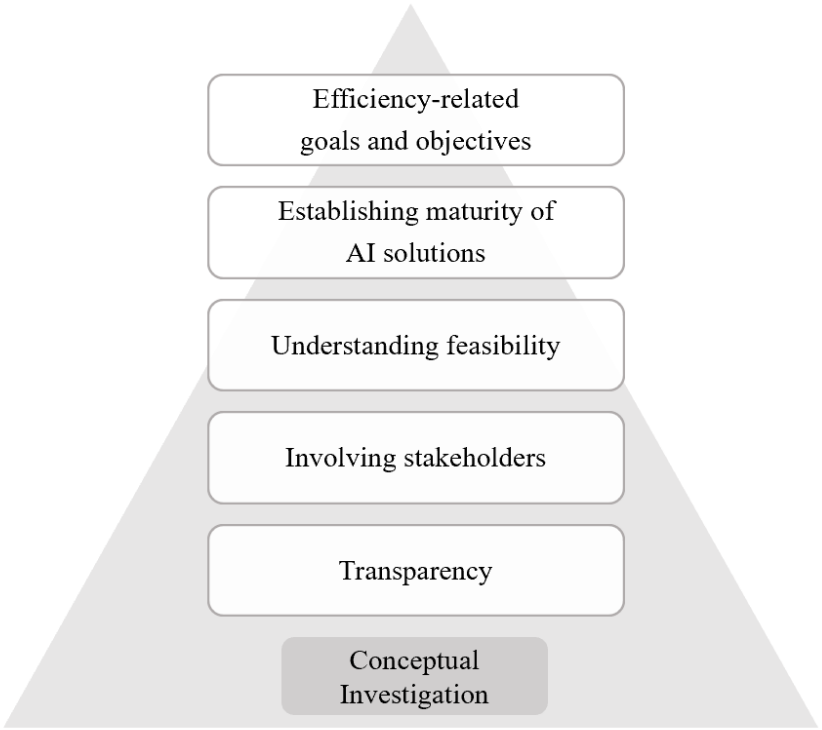

Figure 4. Conceptual investigation themes

In summary, through the conceptual investigation, stakeholders were identified and the impact of the AI solution on these stakeholders explored. Furthermore, this investigation permitted the emergence of values and their sources, which have implications that can be observed in practice. These implications will be further discussed in Section 6.

5.2 Empirical Investigation

The empirical investigation examined the stakeholder's experiences, the context of these experience and understanding of AI. The empirical investigation facilitated in drawing out a number of challenges that arose when translating thematic objectives and values of the AI solution into practical implementation within their organizations.

Data challenges

When asked about the challenges encountered when implementing ethics into practice, R1 shared that the major concerns were more so related to data management than with ethics or moral issues:

> "The fact is that everyone should follow ethics nonetheless. Human rights, they are there for a reason. And I'm always a little bit surprised when people start talking about - we need to consider human rights. And my question is, is there somewhere in Estonia someone who is so far not following human rights or someone who is against human rights?...I don't know any examples."

The "bad" examples of AI, R1 explained, were in no way related to AI. "They have just been bad information system examples...In all other cases, it is a typical rule-based information system. There is nothing to do with AI. I think it is crucial to draw a line that not everything that resembles human intelligence is AI." From R1's experience, the AI solutions cited in some case samples were nowhere near intelligent, referring to rule-based systems being confused with narrow AI. He stated that the black box phenomenon is not universal, and thus it is important to understand what cases truly are out there.

Hard data

Because R2's AI solution did not perform any automated decision-making, there was minimal concern from an ethics point-of-view. "The [human] still has the final word. And we believe the [humans] use it and the outcome [the AI solution] provides is only a support tool." However, data-related concerns were raised by R2.

> "What is one of the problems for example is what we have today, the model of course contains a lot of hard data..." Hard data came in the form of values, numbers, and weighting assigned to intangible human characteristics or traits such as the success and achievement, motivation, intelligence, violence factor, for

> example. Such hard data was used to train the AI solution, whose output would then be validated by a human who doubted its accuracy. Sometimes, the human did not agree with the AI solution's decision or prognosis due to the hard data initially inputted, and vice-versa. Human characteristics were challenging to express in hard, cold data, and at the time of development self-assessed data from users were not available. "But we thought maybe once we have one year's data then we can try already to include in the model."

On the topic of ethics, R2 was aware of the potential bias and discrimination that could occur. However, for their use, they try to find those users who would be most in need of their services and provide them with assistance. To them, this could be seen as as a positive discrimination to identify and direct additional resources for assistance (R2). It was shared that their experience thus far using the AI did not yield any risk of unintentional harm. Lastly, understanding the limitations of the AI solution, R2 recognized that they would not make automated decision-making because the results of the AI solution were not 100% accurate.

Low quality data

R3's AI solution also encountered data-related challenges in the wake of its development. The data used to train the AI prototype had quality issues. Historically, the quality of the data was much lower than at present. Consequently, the AI solution had to spend additional time processing the data to provide an output. In addition, the AI solution required a service in the Estonian language, which was not readily available like English or Russian and thus was more expensive. As a potential user of the tool, R3A was underwhelmed by the AI solution's performance after doing table-top testing of their prototype:

> "They put a check that it is likely to work, but as I said it wasn't so good that I was like "oh wow". We made about 4 test calls...and as I remember, the second test was this when they got the check that it is likely to work. But it took so much time. We already had 3 or 4 seconds to wait that the AI could recognize it…"

Any few seconds saved was crucial in delivering R3's service to the public. A nationally established law dictated a specific timeframe for delivering the service, with which they were obligated to comply. According to R3A, additional improvements were needed and after, there was a possibility to take the AI in use. When asked about the cost-benefit

tradeoff, R3B expressed preference for the benefit ensuring the public's safety and well-being:

> "...if we can save one or two persons with this then it's already paid off. We don't operate exactly like a private company. Sometimes our costs are very high, but this nonmaterial cost, benefits are higher...I think because we are working for [the government], which is there for helping people...We do this for public..."

One of the desirable functionalities of R5's AI solution would have been to have the ability to forecast based on real-time input of data. However, at this stage of development, feasibility and delivering the minimum functionality were important. R5 shared that after analyzing the data, it was clear that at the moment, development of such a feature was not immediately feasible. Once the prototypes proved to be capable, enhancing performance could then be taken up in the future.

Data usability

Since R4's AI solution had undergone the process of advising and funding from R1's organization, consideration for ethical AI frameworks, regulations, and standards were taken into account. R4's organization is heavily regulated. In compliance with regulations, there are terms that state how they can collect, use, process and hold data. As a result, the focus was "strongly on the customer side, how you can use the data and in which way, and how you can process." R4 explained that "it's not just checking how to use and develop, but how to use our data is the main thing we have to consider when doing any project."

It is worth pointing out that R4 as a stakeholder was knowledgeable of the pertinent regulations related to data. In prior experience, R4 developed reports on AI in the public sector and thus had additional background on the topic. Other direct stakeholders involved were also experienced with the GDPR and data collection.

Because of these regulations, R4 pointed challenges related to the usability of the data for the AI solution:

> "We have a lot of data and we had data maybe starting in the 1990s - we have 20 years or more of data we can use. But one or the other reason, you cannot use the data for the purpose of the AI. It's because you may have collected at different times, and different data. For example, there's much more data collected automatically than before. And it's possible that some data can give pretty strong

> signals but you've collected this for only two years. So you're not able to use the data before. So yeah, it's related to cleaning but also with collecting and the reason is we collect them, sometimes data that comes from one excel but you're only using one row, you haven't used the other row, and you're not familiar how to use the other rows, or use it in a very narrow way. And you want to use it differently. So there are lots of problems that's related to that."

When posed the question about the ways in which potential harm was considered, R4 explained that the risk was perceived minimal so long as the AI solution is controlled: "We have to control it, once this machine learns, how it learns and so we can make the correct decisions every time we control it so it has no effect on that side negatively." In terms of potential harm caused by the inaccuracy of AI outputs and decisions, R4 reflected that "we don't have that much deep AI at the moment. But controlling side, I see this as big."

Data management

The data gathered to build R5's AI prototype originated from different data sources and other agencies. Given this, R5 experienced data challenges familiar to other organizations. A considerable amount of time and effort was spent on data management: data clean-ups, formatting, merging, analysis, and so forth. R5 shared that:

> "The data collection methods and formats were very different within the data. So it's actually quite a lot of time in this project we had to dedicate to the data, cleaning steps and data analysis and data merging because this is the thing that was very difficult in this project."

Where other organizations dealt with strict regulatory demands about processing personally identifiable data, with R5's prototype this was not the main concern because the large volumes of data they used did not contain this. However, when asked to discuss ethical frameworks, regulations or issues to possible infringement of individual rights, R5 countered that since no personal data was used, this risk was eliminated.

Nonetheless, the most pressing issue for R5's AI solution was analyzing whether the output of the prototype is reliable. "So I have to analyze why this output from this prototype is something we could believe" (R5). Trust in the accuracy and dependability of the AI solution's output was a contributing reason for why the results of the prototype were not yet communicated with the broader audience or used in daily administrative processes. R5 thought that "it's too early to communicate these because we are not very

sure about the results." Once they establish that the results are indeed reliable, decisions could then be made about releasing the AI solution publicly. Until then, the prototype remained for internal use.

The overall impact of the AI solution could not be fully considered because of the maturity of the tool. R5 explained that the impact assessment cannot be described within the boundaries of the prototype because it is "so minor", much less its impact on society. R5 acknowledged that the broader implications of the AI solution were not assessed in the project because it was a prototype. For them, the societal impact was not an issue because they were in the proof-of-concept stage and it would be too early to describe the impact. This perspective revealed a correlation between the maturity of the AI solution and the consideration for ethical implications of the AI.

The lack of specific data for training was an immediate challenge for certain components of R6's AI solution not to mention navigating through regulatory complexity. Because the AI solution was intended to help the organization comply with the law, there was extensive coverage of the regulatory guidelines that needed adaptations to the solution's architecture.

As an AI vendor, R6 was aware of the ethical frameworks circulated by the EU Council. In addition, R6 was actively involved in different roundtables regarding AI ethics and policies. In terms of ethical impact, R6 stated that it was too soon to foresee the potential impact of the AI solution on stakeholders because the solution was only in development and was not implemented (R6).

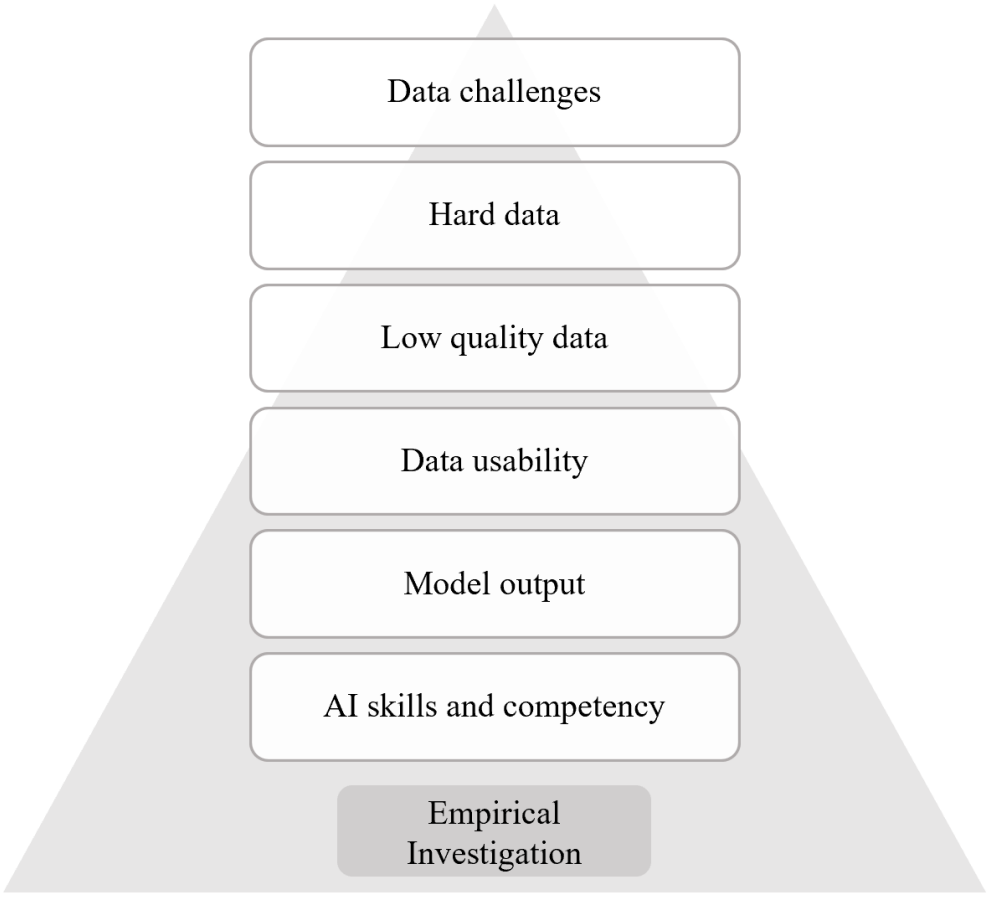

Figure 5. Empirical investigation themes

Model output

As stated previously, R7's AI solution leveraged a facial recognition model that would allow individuals to submit images. These images would then be matched by the AI with publicly available records, the results of which would be sent automatically to the requestor via email.

The facial recognition component had not been trained with custom data from the organization but used with pre-trained public data. However, the programmer tested the model to determine the extent of its functionality using the organization's data.

R7 shared that "the results were not too good, but good and interesting enough" that the organization produced a practical implementation of the AI solution. However, accuracy of the results was a significant area of improvement. R7 shared that the results " are accurate 1 out of 10," attributing this inaccuracy to the fact that the model was "free," "public," and "not top of the line in any way."

R7 observed that even with this low degree of accuracy, a few of their users "had fun with it" while others "had positive results." R7 admitted that sometimes the results befuddled them: "I mean yeah, the computer does what it does and sometimes it is not exact and sometimes I don't understand at all why some faces get connected" (R7). When pressed on whether the programmer would be able to explain, R7 shared that the programmer had a better understanding of how the model works instinctively, but that their technical knowledge of AI could be improved to be able to articulate the *how*.

According to R7, users' feedback had been "mixed." R7 explained that this depended on the expectations users had going into it:

> "We've talked about it publicly, I've stated that the results aren't great, but it's just an interesting thing to test. So if people realize that they will, they have been kind of just interested to see and then positive about the results. But the other half voices that people don't understand what's the point of it, and they didn't get anything and so on. So there are both sides."

In terms of ethical considerations, R7 was somewhat aware of frameworks available after performing high-level research, however, these were not discussed extensively with the programmer. R7 indicated that there was not any time dedicated to think and discuss the ethics "systematically." R7 pointed out that because this project served practical and entertaining purposes, they do not "force" anything on their users, and that users visiting the site use the model out of their own accord. To their knowledge, there were no glaring issues associated with the ethics of AI.

AI skills and competency

To some extent, building the AI solutions through the organization's own expertise became a challenge that organizations sought external assistance. This assistance came in the form of a third-party AI vendor and AI advisors. After the third round in a rigorous tender process, R3B was able to find a vendor that had the skills, competency and authorization to assist in developing the AI solution. The vendor helped the organization refine requirements of the AI solution and produce a prototype. R3B mentioned that although they may not have had the skills to build the prototype, it was a "one-and-done" project and that after, they were able to run the services for the prototype on their own. R3B added that should they require additional competence or assistance, then they would be able to acquire this help from a vendor.

For R5, the novelty of using an AI solution for the first time was accompanied by a steep learning curve. Their AI solution was the first-of-a-kind project in their organization, and as a result, they had to learn quickly. Their public procurement specialist had very limited experience in the past of purchasing a hybrid of what they were used to seeing, which was either IT or market research, but not both as was usually the case with AI. With this project, their specialist had to learn how to procure AI and data analytics-type technology. R5 noted that should they need to buy or procure a similar technology, then they would be better educated and more experienced for the future.

In terms of further developments, the third-party vendor handed over all ownership of the prototype to the organization, including documentation, scripts, codes, and algorithms. Because the prototype was not perceived as a "black-box," they could see what was inside and their IT specialists could take over its maintenance and oversight. However, R5 explained that they currently do not have the ability to further develop the prototype themselves should they want to.

The responses inadvertently underlined the appetite for increasing technical AI competencies and skills. R5 suggested raising the level of competence in AI in general first before thinking about implementing AI solutions. In another situation, R7's AI developer and programmer took initiative by dabbling with free, open source, and publicly available AI models to jumpstart this learning.

As the head agency for AI-related initiatives, R1 provided extensive knowledge, training, and workshops to level the learning curve in AI. In 2019 alone, they offered over 80 different lectures and workshops related to data. Their goal, R1 stated, was to "try and help and support as much as possible." Given that they disbursed AI-related funds, they validated project initiatives by other organizations for success and value-added benefit to citizens and government. Their involvement as advisors also ensured that funding-specific guidelines were respected and strict regulatory requirements related to the processing of personal data. Organizations such as R3A, R3B, R4, R5 and R6 substantially benefited from the guidance and support provided by R1.

The empirical investigation illuminated the challenges that the organizations faced when developing the AI. Data-related issues consistently appeared as the main challenge for organizations when designing and developing their AI solutions. Consequently, ethical concerns were somewhat considered by stakeholders who were more aware of its implications. But because their AI solutions ranged from proof of concept to prototypes, assessing the full breadth of impact of these were deemed premature.

5.3 Technical Investigation

The technical investigation explored the technological mechanisms or features implemented in the AI solutions to discern how these may promote or hinder certain values related to AI ethics. In this section, the following themes were observed.

Privacy

Preserving privacy persisted across all AI solutions. Whenever personal data was involved, R1 assured that all AI projects are reviewed by the data protection agency or that measures are taken to ensure privacy and that projects remain compliant. R1 stated that they ensure citizen privacy is protected, even though their AI solutions do not typically perform citizen-centric analysis, that is to say that their analyses are for general-purpose. Concern for personal data was discussed by R6 who mentioned that one of the most important factors to consider when developing AI for the public sector was the special attention needed for "all aspects related to the processing of big data and personal data."

For R2's AI solution, all requirements related to personal data were considered. For example, training data for the AI solution was anonymized well before it was handed over to researchers. Furthermore, as required by GDPR, they had to assess the impact of personal data. The assessment showed that they were in compliance so long as they use a secure data exchange, anonymize the data and that no automated decision-making was taking place. Due to the nature of the personal data processed by R4, their organization consulted with legal counsel to understand the GDPR's data processing requirements as well. R5 also did not use any personal data because "everything was anonymized." Although the data used was not anonymized, R6 leveraged publicly available data, which did not contain any personal identifiable features, for training the AI solution.

R3A reflected that through a contractual mechanism, it was agreed upon by both parties that the third-party vendor would not "leak" any personal data - intentionally or not. In addition, as part of the terms and conditions of engaging with their third-party vendor, all data must be deleted, including the training data and results. A similar approach was taken by R7's AI solution. In order for their users to feel more secure, R7's AI solution did not retain any of the images or data submitted to the model. As soon as the results were emailed to the user, all data would be erased.

All the organizations exhibited a level of understanding and sensitivity related to handling personal data. Compliance with data protection regulations such as the GDPR was a point of convergence. Where personal data was involved, special attention was given to the law and how this would affect AI projects.

Security

The existence of in certain cases personal data became a precondition for securing the AI solution itself. R2 stated that because there was personal data obtained from data sources such as public registers, general security applied through X-Road. X-Road is Estonia's secured, centrally-managed distributed data-exchange layer. As a result, all data exchanged through X-Road was secured. In addition, the AI solution was developed as an "in-house" tool, meaning that the data processed and outputted by the solution did not go beyond the organization.

A similar case was observed with R3's solution. Although not obtained from other public registers via X-Road, all the data, including sensitive personal data, sat within R3's data centers, protected by firewalls, access and security controls. Being highly regulated, R4 protected and secured its data for the AI solution the same way it would for other data already held by the organization.

Other measures were taken to secure the AI solution and its data such as controlling and restricting access. For example, a password and login combination was required to access the AI solution by those internal to R5's team. And because their AI solution was a prototype, it was housed within the secure servers of the organization. Access was limited. R6 also featured similar security mechanisms. Access was restricted to those with credentials that could be authenticated by the official authentication service managed by the government. The original data, model assessments, and the AI solution itself were hosted on government premises.

R7 did not require users to have registered accounts, only an active email address to which results would be sent. Given that any data submitted to the AI solution would be deleted after processing, the risk of data compromise through theft was minimized. However, access to modifying the AI solution's programming logic level required a central account provisioned to internal employees that could then be recognized and authenticated by the organization.

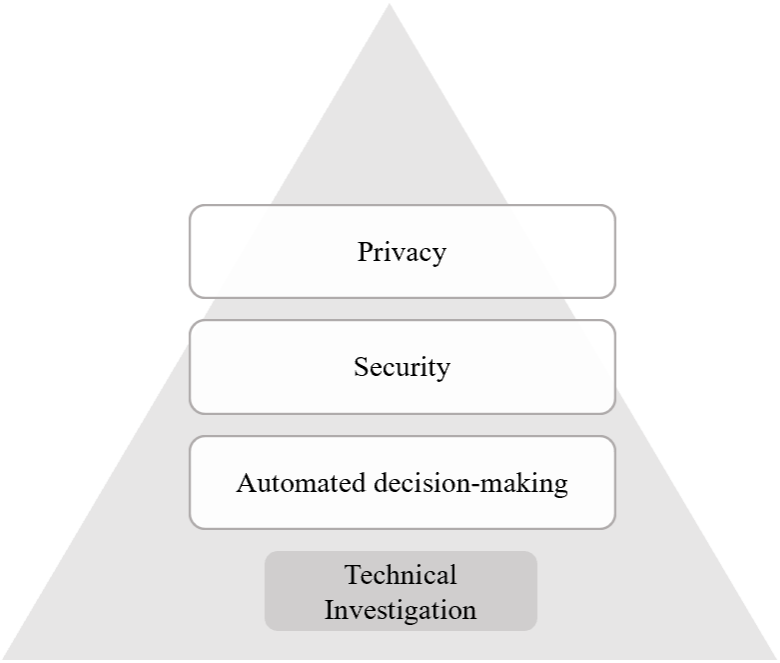

Figure 6. Technical investigation themes

Automated decision making

Although some of the AI solutions discussed thus far had the ability to make decisions, most organizations purposely ended any such automated decision-making with human review, oversight and intervention. Aside from automating manual, routine tasks, R1 mentioned that there was not any automatic decision-making that happened. “There is still some kind of human element in some instances” (R1).

Due to the complex nature of the public services being rendered by R2’s organization, procedures in their purview and the expertise required to perform these services, the output of the process could not be completely automated. Thus, the AI solution served as a decision-support tool that complemented human’s expertise. Because the output of the AI would serve to add to the knowledge of the human to make informed decisions, R2 expressed that the decision-making itself was an element they understood should never be fully automated: “this is something we are aware of. It’s never going to be 100% accurate. And there’s always this human touch that it needs.”

R4 reinforced this view by explaining that certain laws and guidelines advise against automatic decision-making by such tools, referring to the GDPR. R4 also added that due to technical limitations of their AI’s capability and maturity, the automatic decision-making could not be achieved to the same extent a human would have done: “this AI is

not so well-developed that we can talk about deep AI. This is a narrow AI that we are using" (R4). R3B echoed this perspective when discussing the automated decision-making capability of their AI solution: "it is not so advanced yet that the AI can make it by itself."

R6's AI solution also included a human review of all the output. The assessment of the AI's output would still be reviewed and confirmed by an employee of the organization. In addition, R6 added a layer of monitoring by testing each component of the AI solution whenever a change is applied.

Of all the AI solutions in this research, R7's AI model had been the only instance that permitted automated decision-making without human involvement. The matching done by R7's facial recognition model would be performed automatically, and the results would also be returned automatically to the requestor. The output of the model would not be subject to human review; thus, no additional validation would be performed on the matched images.

6 Discussion and Implications

The goal of this research is to grasp the extent in which ethical AI is put into practice among public service organizations. Undertaking this question required an understanding of the issues that public organizations faced in the design and development of AI solutions. Moreover, it also required understanding of how and which ethical principles, if any, were taken into consideration during the design and development of the AI solutions. This section critically discusses the findings that were uncovered by using the VSD tripartite methodology.

6.1 AI design and development challenges

The VSD analysis reveals that a primary value driver for the development of AI in the Estonian public administration is the aim of achieving efficiency and effectiveness in public services through the data analytical means. Multiple respondents point to the desire to perform services more efficiently or to reach a targeted group of the population that required additional attention and resources. This finding confirms extant literature, which had highlighted to the potential of AI in improving the quality of delivering public services as well as cost-savings for the government (Misuraca, et al., 2020; Abbas et al., 2019; Chen et al., 2019; DeSouza et al., 2019). Governments benefit from the efficiency that AI promises to deliver. However, it is the reaping of the benefits associated with efficiency that presents a challenge to governments as they tackle issues related to data management and, consequently, developing and maturing the application of AI solutions.

Public service organizations are beneficiaries of a wealth of data collected over a long period of time, further enriched by exchange and sharing among data sources in different domains. The Estonian public administration, in particular, has enjoyed the richness and diversity of this data thanks to its digitized infrastructure and its well-connected secured, data-exchange layer X-Road, allowing public registers to be shared with other organizations. The challenge then becomes sifting through data that can be used for the purpose of the AI. This issue manifested in a number of different ways, but mainly through quality, usability, and regulatory demands, thus answering the first question of this research:

> *What are the key issues that public service organizations face in AI design and development?*

In an ideal scenario, data collected for the purpose of AI development would come in a structured, compatible, high-quality, machine and human intelligible format, efficiently optimized for processing and training AI. The reality of the situation is often the opposite.

Introducing data with issues or of low quality to AI systems can lead to risks associated with inaccurate, or in some cases, biased outputs (Sousa et al, 2019). Not only that, but low- quality data also affects computing performance of the AI as was the case for one organization. Conversely the AI could potentially require higher computing resources to process.

As a result, a considerable amount of time and effort is dwindled away by the preparation of data through clean-ups, re-formatting, and merging. Janssen et al noted that this tedious task is given less consideration due to the time it takes (2020). High-quality data also demands the infrastructure, computing power, services, and expertise to store, process and manage. These are resources that are finite for a number of public service organizations, unlike its counterparts in the private sector (Desouza et al., 2020).

Regulations such as the GDPR impose certain conditions under which personal data can be processed by an entity (Smuha, 2019). Data may be readily available, but the conditions for which they can be used are limited in scope by data protection regulations. In addition, deducing information based on available, raw data could run afoul of regulations if consent was not obtained. This view is an example of how some organizations are grappling with the best ways to use existing data. For some organizations, the inability to use certain data for purposes outside of the initial terms can hamper the development of AI solutions that seek to become more efficient and effective in delivering public services. This observation supports Desouza, Dawson and Chenok's research on the challenges of AI for the public sector. However, these challenges can apply to organizations in the private sector that are under GDPR jurisdiction as well.

Lack of suitable data for training components of AI solutions add a layer of complexity to the development process. One perspective demonstrated the challenge of translating intangible human traits such as motivation, violence, intelligence, achievement, and so on, into logic that the AI could understand. In literature, Wirtz et al take this a step further and raise the concern about the compatibility of machine and human value judgment. They describe AI systems learning human values that may diverge from the original value system. As the results indicate, the AI solutions are not far advanced to be able to learn by itself.

Hard data, as was referred to by respondents, could not legitimately quantify and attest to the diversity and range of an individual's characteristics, drivers, motivations, emotions,

and qualities. How can the AI therefore offer an output when such data is incomprehensible or unavailable during initial training? These are factors in which AI solutions may never be able to fully grasp. In literature, Dreyfus, an AI critique, argued that intelligence as an intangible human quality cannot be represented in computational programming (Susser, 2013).

Literature has observed that the public sector generally lags behind the private sector in terms of AI maturity (Mehr, 2017; Berryhill et al., 2018). For the majority of the organizations in the study, the maturity of the AI applications seen were at the early developmental stages. That is to say that the AI solutions existed in the form of proof-of-concepts, prototypes or were in the trial or pilot phases. Crucial to attaining efficiency-related goals is to first understand if that which they are trying to solve using AI is feasible. Careful considerations over resources have led organizations to determine feasibility through these means.

Etscheid stated that efficiency can be achieved for a number of administrative procedures using narrow-AI (2019). Indeed, the application of AI has been very limited, and is often viewed and used in the narrow sense. Performance of the AI prototypes, however, have seen mixed results. Some organizations saw consistency between the outputs of the AI compared to the output of the human. Whereas for some, the output was accurate for only 10% of the time. Others still were unable to verify, particularly those that involve long-term forecasting and thus would require monitoring and assessing over periods of time to determine. The speed in which certain AI solutions performed is one area of improvement following unsatisfactory expectations, but that which could be improved in future iterations once feasibility is established.

The novelty of AI presents a steep learning curve for most organizations taking up AI initiatives. The lack of skills and technical competences in this domain is observed as organizations sought guidance through engagements with third-party vendors specializing in AI technology implementation and specially-appointed advisors.

Third-party vendors provide the technical expertise needed to design and develop AI solutions. Contractual, definite engagements with third-party vendors, often from the private sector, have afforded some organizations the ability to take-up AI initiatives and evaluate the feasibility of AI solutions without needing to invest into the fully-developed solutions upfront. Successful engagements can encourage future developments in organizations. However, procurement of these services proved to be a challenge initially for public servants who lacked general knowledge about AI. Because AI is new to most organizations, public servants are unfamiliar with navigating through the technical requirements and feasibility of building such technologies, the process of procuring

vendors and services took a significant amount of time, not to mention the already rigorous process of vetting vendors against procurement laws. Nonetheless, third-party vendors have a degree of influence over the outcomes of AI projects, serving as consultants, advisors, and implementers of AI solutions for public service organizations.

In addition, there is a demonstrated appetite for acquiring AI- and data-related skills by a range of public servants, from business analysts to programmers. Self-initiated learning has developed some public servants' AI-skills, while supportive management has enabled application of these self-acquired skills in practice. Perhaps one contributing factor to the success of some AI-initiatives is the overarching support provided by the government-appointed agency for AI-initiatives. Advisors from this agency offered substantial guidance throughout by advising, training, and funding.

6.2 Consideration of AI ethical principles

It has been deemed important to first discuss the challenges that AI presented to the public service organizations because results show that these challenges superseded concerns for the ethics of AI. At first, the results conveyed little to no consideration for the ethics of AI by public service organizations, owing to the immaturity of the AI solutions and minimal application of ethical AI frameworks. However, the VSD approach has been instrumental in uncovering the different ways in which ethical principles were considered in the design and development of the AI solutions by public service organizations. The principles in action may not be obvious or concrete, however, they were activated to a certain degree; some principles more operationalized than others. Thus, the following subsections address the second research question and are organized according to AI4People Ethical Framework:

> *In what ways are AI ethical principles considered in practice by public service organizations in the design and development of AI for public service delivery?*

Beneficence

The conceptual investigation imparted the goals of public service organization's aspirations for using AI solutions. This investigation showed that efficiency and effectiveness in order to improve delivery of public services were the main values at play. Although not an ethical principle in of itself, the intent was to deliver better quality services for the benefit of the citizens being served. Beneficence is the promotion of good through alignment with human values, the prioritization of human well-being in the design of systems, and that AI should serve to benefit humanity and common good (Floridi & Cowls, 2019). As public service institutions, these organizations are held up to a set of public values, one of which is to serve or contribute to the common good (Jørgensen & Bozeman, 2007).

Non-maleficence

This principle is manifested in tangible measures taken to ensure privacy, security, and safety. The technical investigation expounded on these measures. The specific handling of personal data and how this is protected by security mechanisms attest to this principle. Authentication by means of passwords, secured servers and data exchange, and protecting the AI solution within closed systems with strict access controls were demonstrated by

public service organizations. Although not in service of AI ethical principles per se, these practices are a by-product of stringent regulations such as the GDPR requiring such measures.

Without the existence of personal data, there is less risk posed to privacy and security. Thus, some organizations use the approach of intentionally removing personal data from sets or obfuscating this information through anonymization techniques to allow its use for the AI solution. This method reduces the concern for privacy.

While adherence to ethical guidelines may not be the driving factor for this principles, public service organizations are, indirectly, abiding by this principle. The findings here align with empirical research conducted by Ryan et al. in which security and privacy ethical concerns are addressed due in part to the short-term impact and relative proximity of these as a result of legal obligations. Thus, ethical principle of non-maleficence through security and privacy become considered in this way.

Autonomy

In the context of autonomous AI, human choice is central to this principle (Floridi et al., 2018). As observed in practice, the AI solutions are not so advanced to perform automatic decision-making by itself. Nonetheless, in cases where automated decision-making would occur, reviews of the AI's output are done by the human, and the final decision resides with the human.

Furthermore, a number of public servants are more sensitive to the risks involved with automated decision-making, which is a significant concern in AI ethics, but this awareness has the propensity to stem from data-related regulations, specifically GDPR's Article 22. The weight of the law and its repercussions - a hefty fine levied on noncompliant organizations or reputational damage (General Data Protection Regulation, 2016) - fall heavily on the shoulders of organizations processing sensitive data. As a result, extra attention is paid to AI activities involving the use of personal data.

Justice

The stakeholders involved in the design and development of the AI solutions have been, for the most part, limited to key, direct stakeholders. That is to say, these direct stakeholders often are small teams composed of people attentive to ensuring the functionality of the AI. Indirect stakeholders, those who may not necessarily use the AI but are implicated by its use, have not been consistently involved in these early stages. Indirect stakeholders in most cases are the citizens who receive or benefit from the public services delivered. A lack of diversity may affect the way stakeholder values are represented and influence the design of the AI.

Owing to the difficulty in translating intangible, lofty concepts such as human motivation, achievement, violence into digestible machine logic, hard data has challenged some organizations to rethink the reliability of the AI's output. Add to that the inconsistency between the human's assessment and the AI's output, these concerns instill doubt on the reliability of the AI's solution.

Researchers describe a pattern in which when errors occur within automated systems, trust in the system is reduced and is carried over to similar systems (Keziemski & Misuraca, 2020). However, this context is different and relates to the quality of data used instead. "Can we trust the output of the AI?" Users of decision-support tools are aware that this could affect indirect stakeholders such as the citizens who are impacted by such decisions. Hence, organizations place humans to decide ultimately, a kind of guardian for these decisions before they act on such information.

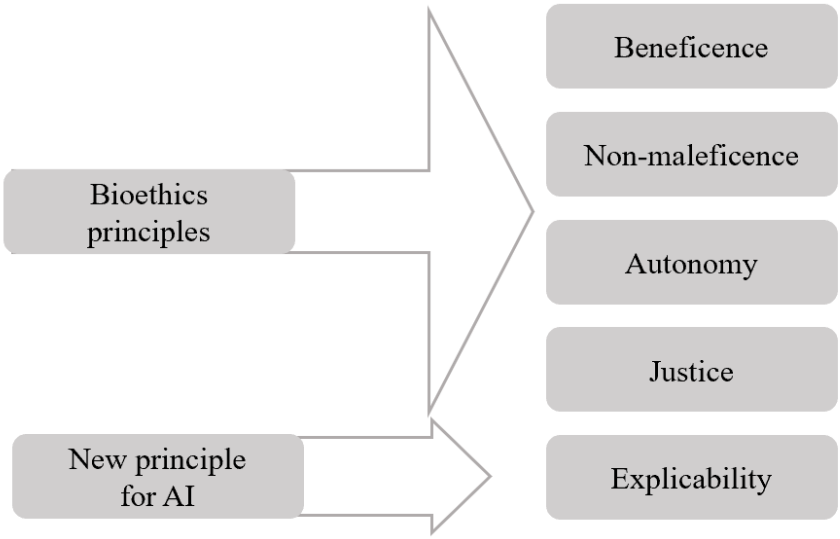

Figure 7. AI4People ethical principles (2018)

Furthermore, public servants using AI for delivering services that affect the well-being of certain groups of populations are highly conscious about the bias that may exist when targeting these groups. However, with or without the AI or similar tools, public servants would still have to perform tasks that may be deemed non-inclusive or even discriminatory, but this is the nature of the work. The services are available to everyone who needs them. AI is seen as a way to provide these services more efficiently. Floridi et al suggest that the principle of justice refers to using AI to correct past wrongs, to ensure that it creates shared benefits and prevent new harms in current social structures. Other researchers suggest otherwise, that AI or other technologies should not be the focus of this debate on cultural values. Rather, it helps society have these discussions on the kinds of cultural values it prioritizes when faced with social or ethical issues and if the citizens are in agreement with this (Bryson & Kime, 1998). The cause for real concern is the misuse of the AI, which as observed, is not the case as no harm has been reported to date.

Explicability

One aspect of this principle observed in practice is the ability to explain the AI and how it arrived at its decision. The results of the interviews indicate that the “black-box” phenomenon is not prevalent. Instead, narrow AI is observed to improve administrative procedures that could otherwise be performed in Excel albeit with more time and effort. To this extent, explaining the “how” behind the AI is not a concern. In addition, third-party vendors who may have initially built the AI solution hand over documentation, scripts, codes and algorithms to the organizations, and their IT specialists are able to continue with maintenance of the solution. Procurement contracts specify these terms, and solutions are not proprietary.

Another aspect of the principle of explicability relates to transparency. The results conveyed that all the respondents, who are public servants, seem to favor informing the public of the use of AI in the delivery of public services. However, this is not yet done in practice due to the immaturity of their AI solutions and that they are not currently being used, except for two cases where one is already deployed for public use and the other being piloted by the organizations. The first case had been transparent about an AI delivering the service. The other case had wrestled with this dilemma between transparency and beneficence, or preserving the well-being of citizens.

On one hand, informing those receiving the service about the involvement of an AI is an act of transparency. On the other hand, delivering this information, particularly when the decision is negative, could affect the well-being of the citizen. Thus, here values of

transparency and beneficence conflict. In addition, this dilemma raises the question: who should be trusted more for accuracy? The AI that administers decisions based on hard, cold data or the compassionate human who may not have all the information? This conundrum reflects the way in which Bryson et al suggest that AI can help society understand its own values (1998).

The high regard for transparency by public servants comes in stark contrast with the views in the private sector. Whereas public servants feel the need to inform the public about their use of AI in processes, the private companies are not so obliged. Ryan et al's empirical research focused on the private sector, wherein a number of companies interpret transparency in terms of systems and algorithms, and much less on transparency with public authorities (2021). Unless legally obligated, the private sector may not act on this on purpose as it could hinder them from an economic and financial advantage.

This difference shows a divergence in values between the public and private sector which could have implications for achieving values alignment in future policymaking, as pointed out in literature by Smuha (2019). Resolution of these differences should be addressed in order to provide a clearer path for organizations to further their AI application.

The theme of early developmental stage correlates with the level of consideration relegated to the ethics of AI. The concern for the ethics of AI is overshadowed by much more pressing, immediate data challenges. Organizations are focused on establishing feasibility of the AI. But because the AI solutions are in such an early stage of development, some not even in use by their teams or released beyond organizations, the concern for AI risk and ethics is significantly diminished as a result. This finding affirms Ryan et al findings on the temporality perceived on ethical issues by practitioners (2021).

Furthermore, automated decision-making features of AI, which are a cause for concern, are not a reality for a number of the AI solutions due to their "narrow" application and that a human still performs review and ultimately decides.

It could be suggested insofar that the less mature an AI solution is, the less risk exists for unintentional harm. Following this logic, it could be argued that since some of these applications of AI are termed narrow, the unintentional harm that can arise from these applications are minimal to none. Simply put, it is far too soon to say because the applications are not fully developed nor advanced to cause harm yet.

There is some degree of awareness by public servants on the risks that are posed by AI. Some public servants were more aware than others due to their exposure to the topic of

AI in general or having had previous work experience with data analytics. Others obtained knowledge through workshops hosted by AI advisors.

As observed through the head AI-agency, data privacy impact assessments, risks assessments, data management protocols and standards were shown to be a component of a checklist created to comply with regulations. These tools help assess an AI project's qualification for receiving funding, and as a result, these projects are better informed about the legal obligations regarding processing personal data. However, ethical AI guidelines, standards, or frameworks were minimally consulted by public servants involved in the design and development of the AI. These instruments were seen more as a project's tick-box activity than as a substantive requirement that affects the AI throughout the project's evolution.

Taking all into consideration, these results shed light to the main research question, which is:

> *How do public service organizations ensure ethically-aligned AI public services in practice?*

Public service organizations design and develop AI that are aligned with the intent of improving public services for the benefit of public good. Values of efficiency and effectiveness are the main driver to achieve this intention. To some extent, AI ethical principles of beneficence, non-maleficence, justice, explicability are indirectly considered and are somewhat practically demonstrated in myriad of ways including: compliance with privacy and data protection regulations; the development of AI solutions with built-in security measures to protect data and privacy; a degree of awareness of the potential inaccuracy of the AI's output and how this may discriminate against certain groups or affect indirect stakeholders; and discretion for transparency when using AI to deliver public services to society. Thus, in this way, ethical AI is put into practice, however, less rigorously and systemically due to challenges associated with data, AI skills and competencies, and the immaturity of AI development in general.

In literature, Siau and Wang suggest that in order to build ethical AI, there should be an understanding of the ethics of AI and recommend placing AI ethics at the center when developing AI and not only after the development of the AI (2020). However, it becomes evident, therefore, that "soft law," or the ethics side does not carry much weight in the design and development of AI by public service organizations the same way hard laws have done. Compliance with legislation such as the GDPR engenders more attention to the risks posed by AI. Jobin et al suggest that AI ethical codes and laws should become aligned so that the global community can move forward towards an ethically designed AI (2019, p. 396). Indeed, the recent proposal for an AI regulation by the European Commission is regarded as a positive foot in the right direction by many.

6.3 Implications

Following this discussion, the findings have some practical implications for designers, governments, and policymakers:

Designers of AI solutions should actively consider principles early in the design phase and throughout the development phase. In addition, indirect stakeholders such as citizens should also be involved in the design of AI systems that deliver public services or interact with the public as they are implicated by their use. Indirect stakeholder input could help address value conflicts and design AI solutions that are aligned with ethical values.

Governments should continue develop a rich data ecosystem that enables sharing and exchange of highly quality data while maintaining security and integrity. Good data management practices should be encouraged as this can increase the uptake of AI initiatives. In addition, resources should be provided to increase competence and skills in the AI domain. Initiatives that encourage AI uptake whether through data sharing, funding, training and public events can bolster AI knowledge. Engagements with third-party AI vendors from the private sphere tend to generally have expertise and knowledge, which can be beneficial for spurring innovation. Viewed as the technical experts, third-party AI vendors are in a valuable position to bolster awareness and implementation of the ethics of AI.

The application of AI in the public sector is at its infancy, while regulation of AI is on the horizon. Thus, there are opportunities as well as risks. Progress in regulatory space can provide clear guidance and direction in standardizing ethical principles and operationalizing them. Policymakers should examine the impact of proposed AI regulations on innovation as they could hamper them. At the same time, policymakers should continue working with agility to calibrate legislation based on-the-ground input from all stakeholders and validate with empirical data.

6.4 Limitations

In consideration of the limitations of this research, the applicability of these findings are impacted by the context in which they were formulated. First, the cases selected are based in the European context, in particular Estonia. Thus, some findings may not be applicable to other countries or regions. Secondly, the set list of ethical principles was informed by literature, but not obtained from an empirical, bottom-up approach. The ethical principles discussed have reached a level of consensus but are in no means universal. Nonetheless, these findings reflect the conditions occurring in practice in the Estonian public sector and contribute to wider discussions on AI ethics.

7 Conclusion

The application of AI is growing and affecting aspects of society, both in the private and public spheres. Along with the opportunities of AI are the risks of exacerbating societal ills, infringing on privacy, and loss of human choice. In an attempt to abate these risks, institutions and academics have stimulated the discussions on the ethics of AI, producing ethical frameworks, standards and even moving to regulating the field.

This research specifically takes on the topic of AI ethics by juxtaposing these ethical concerns and the actual implementation of AI ethics in the public sector. More precisely, this research offered insights into how public service organizations are ensuring that ethical values are aligned and translated in the design and development of AI for the delivery of public services.

Using the Value Sensitive Design as a theoretical and methodological approach, the results of this research indicate that ethics of AI is being considered to a certain degree. Public service organizations indirectly translate ethical principles by way of addressing requirements for AI's functionality and requirements imposed by regulations such as the GDPR. However, the maturity of AI solutions is in such early stages of development that systematic consideration for and application of AI ethical principles is overshadowed by more pressing, practical issues related to the feasibility of AI solutions and data management.

Furthermore, a level of awareness exists among the public servants for the risks posed by AI. Their knowledge, skills and competences in general AI can be raised through AI training initiatives. Where third-party AI vendors play a role in bridging this skills gap through procurement, they are also in a position to serve as both technical and ethical advisors to public service organizations seeking their guidance in the design and development of AI.

These research findings fill a gap in sparse empirical-based scholarship on the ethics of AI. However, they are by no means sufficient to address the continuous debates on "what values" and "whose values" in ethical AI development. Therefore, suggested future areas of research on AI ethics in the public sector should examine citizen's perception on the use of AI in delivering public services. Another avenue is to explore whether certain public sector values conflict with AI ethical principles, as well as how AI is inadvertently supporting values such as corruption. These areas of further research are some additional steps that can be taken towards advancing the dialogue on AI ethics in an ever-evolving, culturally complex society and building a conscionable future for generations to come.

References

Abbas, N. N., Ahmed, T., Shah, S. H. U., Omar, M., & Park, H. W. (2019). Investigating the applications of artificial intelligence in cyber security. In *Scientometrics* (Vol. 121, Issue 2, pp. 1189–1211). https://doi.org/10.1007/s11192-019-03222-9

Adams, S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., Hall, J. S., Samsonovich, A., Scheutz, M., Schlesinger, M., Shapiro, S. C., & Sowa, J. (2012). Mapping the Landscape of Human-Level Artificial General Intelligence. *AI Magazine*, *33*(1), 25–42. https://doi.org/10.1609/aimag.v33i1.2322

Aoki, N. (2020). An experimental study of public trust in AI chatbots in the public sector. *Government Information Quarterly*, *37*(4), 101490. https://doi.org/10.1016/j.giq.2020.101490

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Barth, T. J., & Arnold, E. (1999). Artificial Intelligence and Administrative Discretion: Implications for Public Administration. In *American Review of Public Administration* (Vol. 29, Issue 4, pp. 332–351).

Bartneck, C., Lütge, C., Wagner, A., & Welsh, S. (2021). What Is Ethics? In C. Bartneck, C. Lütge, A. Wagner, & S. Welsh (Eds.), *An Introduction to Ethics in Robotics and AI* (pp. 17–26). Springer International Publishing. https://doi.org/10.1007/978-3-030-51110-4_3

Behar, J. E. (1993). Critique of Computer Ethics: Technology as Ideology. *Journal of Information Ethics*, *2*(2), 27-43,95.

Berryhill, J., Heang, K. K., Clogher, R., & McBride, K. (2019). *Hello, World: Artificial intelligence and its use in the public sector*. https://doi.org/10.1787/726fd39d-en

Borning, A., & Muller, M. (2012). Next steps for value sensitive design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1125–1134). Association for Computing Machinery. https://doi.org/10.1145/2207676.2208560

Bryson, J., & Kime, P. (1998). *Just another artifact: Ethics and the empirical experience of AI* (p. 390).

Chen, T., Guo, W., Gao, X., & Liang, Z. (2020). AI-based self-service technology in public service delivery: User experience and influencing factors. *Government Information Quarterly*, 101520. https://doi.org/10.1016/j.giq.2020.101520

Crevier, D. (1993). *AI: The Tumultuous History of the Search for Artificial Intelligence* (p. 386).

Cummings, M. L. (2006). Integrating ethics in design through the value-sensitive design approach. *Science and Engineering Ethics*, *12*(4), 701–715. Scopus. https://doi.org/10.1007/s11948-006-0065-0

Daly, A., Hagendorff, T., Li, H., Mann, M., Marda, V., Wagner, B., Wang, W. W., & Witteborn, S. (2019). *Artificial Intelligence, Governance and Ethics: Global Perspectives* (SSRN Scholarly Paper ID 3414805). Social Science Research Network. https://doi.org/10.2139/ssrn.3414805

Davenport, T., Brynjolfsson, E., McAfee, A., & Wilson, J. (2019). *Artificial Intelligence: The Insights You Need from Harvard Business Review.* Harvard Business Review Press. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2003692

Davis, J., & Nathan, L. P. (2015). Value Sensitive Design: Applications, Adaptations, and Critiques. In J. van den Hoven, P. E. Vermaas, & I. van de Poel (Eds.), *Handbook of Ethics, Values, and Technological Design: Sources, Theory, Values and Application Domains* (pp. 11–40). Springer Netherlands. https://doi.org/10.1007/978-94-007-6970-0_3

Desouza, K. C., Dawson, G. S., & Chenok, D. (2020). Designing, developing, and deploying artificial intelligence systems: Lessons from and for the public sector. *Business Horizons*, *63*(2), 205–213. https://doi.org/10.1016/j.bushor.2019.11.004

Dreyfus, H. L. (1974). Artificial Intelligence. *The ANNALS of the American Academy of Political and Social Science*, *412*(1), 21–33. https://doi.org/10.1177/000271627441200104

Etscheid, J. (2019). *Artificial Intelligence in Public Administration: A Possible Framework for Partial and Full Automation* (pp. 248–261). https://doi.org/10.1007/978-3-030-27325-5_19

European Commission. (2019). *Digital Government Factsheet—Estonia*. https://digital-strategy.ec.europa.eu/en/policies/desi

General Data Protection Regulation, Pub. L. No. (EU) 2016/679 (2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj

EU Proposal AI Regulation, 2021/0106 (COD) (2021). https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, *7*(1), 1–9. https://doi.org/10.1057/s41599-020-0494-4

Floridi, L. (2013). *The Ethics of Information*. OUP Oxford.

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Lütge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*. https://doi.org/10.1007/s11023-018-9482-5

Franklin, S. (2014). History, motivations, and core themes. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 15–33). Cambridge University Press. https://doi.org/10.1017/CBO9781139046855.003

Friedman, B., Kahn, P. H., & Borning, A. (2008). Value Sensitive Design and Information Systems. In *The Handbook of Information and Computer Ethics* (pp. 69–101). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470281819.ch4

Goertzel, B., & Pennachin, C. (Eds.). (2007). *Artificial general intelligence*. Springer.

Gomes de Sousa, W., Pereira de Melo, E. R., De Souza Bermejo, P. H., Sousa Farias, R. A., & Oliveira Gomes, A. (2019). How and where is artificial intelligence in the public sector going? A literature review and research agenda. In *Government Information Quarterly* (Vol. 36, Issue 4, p. 101392 [1-14]). https://doi.org/10.1016/j.giq.2019.07.004

Government of the Republic of Estonia. (2019, July). *Estonia's national artificial intelligence strategy 2019-2021*. Artificial Intelligence for Estonia. https://en.kratid.ee/

Grosz, B., Altman, R., Mackworth, A., Mitchell, T., Horvitz, E., Mulligan, D., & Shoham, Y. (2016). *Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence*. https://ai100.stanford.edu/sites/g/files/sbiybj9861/f/ai_100_report_0831fnl.pdf

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

IEEE. (2019). *IEEE Position Statement Artificial Intelligence*. https://globalpolicy.ieee.org/wp-content/uploads/2019/06/IEEE18029.pdf

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy Artificial Intelligence. In *Government Information Quarterly* (Vol. 37, Issue 3, p. 101493 [1-8]). https://doi.org/10.1016/j.giq.2020.101493

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Jørgensen, T. B., & Bozeman, B. (2007). Public Values: An Inventory. *Administration & Society*, *39*(3), 354–381. https://doi.org/10.1177/0095399707300703

Kerikmae, T., & Parn-Lee, E. (2020). Legal dilemmas of Estonian artificial intelligence strategy: In between of e-society and global race. In *Ai & Society* (p. pre-print). https://doi.org/10.1007/s00146-020-01009-8

Krafft, P. M., Young, M., Katell, M., Huang, K., & Bugingo, G. (2019). Defining AI in Policy versus Practice. *ArXiv:1912.11095 [Physics]*. http://arxiv.org/abs/1912.11095

Legg, S., & Hutter, M. (2007). Universal Intelligence: A Definition of Machine Intelligence. *ArXiv:0712.3329 [Cs]*. http://arxiv.org/abs/0712.3329

Mehr, H. (2017). *Artificial Intelligence for Citizen Services and Government*. 19.

Misuraca, G., van Noordt, C., & Boukli, A. (2020). The use of AI in public services: Results from a preliminary mapping across the EU. In Y. Charalabidis, M. A. Cunha, & D. Sarantis (Eds.), *13th International Conference on Theory and Practice of Electronic Governance (ICEGOV 2020)* (pp. 90–99). Association for Computing Machinery. https://doi.org/10.1145/3428502.3428513

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, *1*(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

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

OECD. (2019). *Recommendation of the Council on Artificial Intelligence*. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Perri 6. (2001). Ethics, Regulation and the New Artificial In[TEL]ligence, Part I: Accountability and Power. *Information, Communication & Society*, *4*(2), 199–229. https://doi.org/10.1080/713768525

Reis, J., Santo, P. E., & Melão, N. (2019). Artificial Intelligence in Government Services: A Systematic Literature Review. In Á. Rocha, H. Adeli, L. P. Reis, & S. Costanzo (Eds.), *New Knowledge in*

Information Systems and Technologies (pp. 241–252). Springer International Publishing. https://doi.org/10.1007/978-3-030-16181-1_23

Robles, G., Gamalielsson, J., & Lundell, B. (2019). *Setting Up Government 3.0 Solutions Based on Open Source Software: The Case of X-Road* (pp. 69–81). https://doi.org/10.1007/978-3-030-27325-5_6

Russell, S. J., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (Fourth edition). Pearson.

Ryan, M., Antoniou, J., Brooks, L., Jiya, T., Macnish, K., & Stahl, B. (2021). Research and Practice of AI Ethics: A Case Study Approach Juxtaposing Academic Discourse with Organisational Reality. *Science and Engineering Ethics*, *27*(2), 16. https://doi.org/10.1007/s11948-021-00293-x

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management*, *31*, 74–87. https://doi.org/10.4018/JDM.2020040105

Simmons, A. B., & Chappell, S. G. (1988). Artificial intelligence-definition and practice. *IEEE Journal of Oceanic Engineering*, *13*(2), 14–42. https://doi.org/10.1109/48.551

Slee, T. (2020). The Incompatible Incentives of Private-Sector AI. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 106–123). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.6

Smuha, N. A. (2021). From a 'race to AI' to a 'race to AI regulation': Regulatory competition for artificial intelligence. *Law, Innovation and Technology*. https://www.tandfonline.com/doi/abs/10.1080/17579961.2021.1898300

Sun, T. Q., & Medaglia, R. (2019). Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare. In *Government Information Quarterly* (Vol. 36, Issue 2, pp. 368–383). https://doi.org/10.1016/j.giq.2018.09.008

Susser, D. (2013). Artificial Intelligence and the Body: Dreyfus, Bickhard, and the Future of AI. In V. C. Müller (Ed.), *Philosophy and Theory of Artificial Intelligence* (pp. 277–287). Springer. https://doi.org/10.1007/978-3-642-31674-6_21

Thierer, A., Sullivan, A., & Russell, R. (2017). *Artificial Intelligence and Public Policy*. https://doi.org/10.13140/RG.2.2.14942.33604

Umbrello, S., & De Bellis, A. F. (2018). *A Value-Sensitive Design Approach to Intelligent Agents* (SSRN Scholarly Paper ID 3105597). Social Science Research Network. https://papers.ssrn.com/abstract=3105597

Umbrello, S., & van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00038-3

UNESCO. (2021). *Draft text of the Recommendation on the Ethics of Artificial Intelligence—UNESCO Digital Library*. https://unesdoc.unesco.org/ark:/48223/pf0000376713.locale=en

Vakkuri, V., Kemell, K., Kultanen, J., & Abrahamsson, P. (2020). The Current State of Industrial Practice in Artificial Intelligence Ethics. *IEEE Software*, *37*(4), 50–57. https://doi.org/10.1109/MS.2020.2985621

van Noordt, C., & Misuraca, G. (2020). Exploratory Insights on Artificial Intelligence for Government in Europe. *Social Science Computer Review*, 0894439320980449. https://doi.org/10.1177/0894439320980449

van Wynsberghe, A. (2013). Designing Robots for Care: Care Centered Value-Sensitive Design. *Science and Engineering Ethics*, *19*(2), 407–433. https://doi.org/10.1007/s11948-011-9343-6

Vries, H. D., Bekkers, V., & Tummers, L. (2016). Innovation in the Public Sector: A Systematic Review and Future Research Agenda. *Public Administration*, *94*(1), 146–166. https://doi.org/10.1111/padm.12209

Wirtz, B. W., & Müller, W. M. (2019). An integrated artificial intelligence framework for public management. *Public Management Review*, *21*(7), 1076–1100. https://doi.org/10.1080/14719037.2018.1549268

Wirtz, B. W., Weyerer, J. C., & Sturm, B. J. (2020). The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration. *International Journal of Public Administration*, *43*(9), 818–829. https://doi.org/10.1080/01900692.2020.1749851

Zhang, Y., Wu, M., Tian, G. Y., Zhang, G., & Lu, J. (2021). Ethics and privacy of artificial intelligence: Understandings from bibliometrics. *Knowledge-Based Systems*, *222*. Scopus. https://doi.org/10.1016/j.knosys.2021.106994

Zhuang, Y., Wu, F., Chen, C., & Pan, Y. (2017). Challenges and opportunities: From big data to knowledge in AI 2.0. *Frontiers of Information Technology & Electronic Engineering*, *18*(1), 3–14. https://doi.org/10.1631/FITEE.1601883

Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, 101577. https://doi.org/10.1016/j.giq.2021.101577

Appendix

A Interview Guide

Background:

1. Can you please introduce yourself and your role in the development of the AI, hereto referred as AI solution?
2. From your perspective, how do you define artificial intelligence (AI)?
3. Can you discuss how this project came about?
    a. How was the head agency involved in this project, if at all?
4. For what purpose or intention was this AI developed?
5. Who were the different stakeholders involved in the design of the AI?
    a. Could you describe briefly the composition of these stakeholders?
6. Does your organization have an AI strategy? If so, how does this align with your organization's ethics or values?

Technical components:

7. Please describe the AI - briefly how it works at a high-level.
8. What data was used to train the AI?
    a. How was the training performed?
    b. Could you describe any privacy-preserving methods used on the data?
9. What is the output of the AI?
    a. If decisions are made independently by the AI, can this be explained by its developers?
    b. What oversight or mechanisms are in place to address potential misjudgments or unintended harm caused by the AI?
10. What security measures are taken to protect the data *and* the AI itself?
11. Please describe the governance approach taken by your organization for the AI.
    a. What kinds of ongoing checks are done to ensure that the AI is functioning as intended?
    b. Where a third-party vendor was involved in the development of the AI, how is this relationship managed by the public service organization?

Design Considerations

12. What informed the design requirements for the AI?
13. In what ways did you involve the stakeholders in the design of the AI?
14. How was the project team made aware of any risks involved with the development of the AI?
15. Was the team aware of any ethical frameworks or regulations related to AI?
16. How was the potential impact of this AI on stakeholders considered?
17. What were some challenges or issues raised with the design of this AI?

User Acceptance:

18. For public services that use the AI, were the end-users aware that an AI was used to deliver the service?
19. What has been the users' feedback on the AI and how was this collected?
20. Were there any challenges with the use of the AI?
21. Has the AI met its intended objectives, if so, in what ways? If not, please describe.

Final Questions

22. With regards to AI in general, what risks do you see in relation to its use in society?
23. What would be the most important factor to consider when developing AI for the public sector?

B Interview respondents

Respondent	Respondent's Role	Data Collection Date	Data Collection Format
R1	Data and AI specialist	05 March 2021	Semi-structured interview
R2	IT service developer	01 April 2021	Semi-structured interview
R3A	Development specialist	28 May 2021 07 June 2021	Written responses followed by a semi-structured interview
R3B	Technical procurement specialist	27 May 2021	Semi-structured interview
R4	Technology development specialist	26 May 2021	Semi-structured interview
R5	Data analyst	02 June 2021	Semi-structured interview
R6	Third party AI developer	27 May 2021	Preferred written-responses
R7	AI project lead	01 April 2021	Semi-structured interview
R8	AI product manager	26 May 2021	Email response

Total number of respondents: 9

Declaration of Authorship

I hereby declare that, to the best of my knowledge and belief, this Master Thesis titled "The State of Ethical AI in Practice: A Multiple Case Study of Estonian Public Service Organizations" is my own work. I confirm that each significant contribution to and quotation in this thesis that originates from the work or works of others is indicated by proper use of citation and references.

Tallinn, 09 August 2021

Charlene Palomer Hinton

Consent Form

for the use of plagiarism detection software to check my thesis

Name: Hinton
Given Name: Charlene Palomer
Student number: 509651
Course of Study: Public Sector Innovation and eGovernance
Address: Schlossplatz 2, 48149 Münster
Title of the thesis: The State of Ethical AI in Practice: A Multiple Case Study of Estonian Public Service Organizations

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

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

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

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

Tallinn, 09/08/2021

Charlene Palomer Hinton