International Journal of Technoethics
Volume 14 • Issue 1

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

Charlene Hinton, Erasmus Mundus PIONEER, Belgium*

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 varies according to the maturity of the AI and the public servant's level of awareness, knowledge, and competences in AI. Data-related challenges persist as public service organizations work on fine-tuning their AI applications.

KEYWORDS

AI Ethical Principles, Artificial Intelligence, Estonia, Ethical AI Design and Development, Ethics, Practical Application of AI Ethics, Public Sector, Value Sensitive Design

1. INTRODUCTION

Artificial intelligence (AI) has a deep potential to change various aspects of citizens' daily lives and of society as a whole. A systematic review of academic literature has shown growth in the uptake of artificial intelligence in the public sector (Gomes de Sousa et al., 2019; Berryhill et al., 2019; van Noordt & Misuraca, 2020). In Europe alone, the use of AI in public services is increasing, with over 230 empirical use cases identified (van Noordt & Misuraca, 2020). Researchers have noted that 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 citizens' privacy and security, loss of decision-making autonomy, and unintentional harm that arise from AI systems may reinforce existing discriminatory practices (Sun & Medaglia, 2019).

DOI: 10.4018/IJT.322017 *Corresponding Author

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 communities, putting into question the human values that are deemed important by society.

Against the background of the 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 about the practical implementation of the 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 a limited understanding of the implications of the use of AI. They hypothesize that insufficient research on empirical, context-based AI usage in governments can induce systemic failures that may negatively impact not only governments but also societies as a whole (Zuiderwijk et al., 2021). Therefore, this research aims to address this knowledge gap in the rapidly-evolving field of AI by addressing the following questions:

1. 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 the design and development of AI?
    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 research contributes to a broader academic discussion about the practical implementation of AI ethics and concurrently maintains focus on the under-researched public sector within the AI discipline. Furthermore, Estonia is chosen as the country context of study given its highly digitalized public services, its aggressive AI strategy, and the extensive collection of use cases of AI in the public sector. The rest of this research is organized as follows: Section 2 offers research background on AI in the public sector and the debates concerning AI ethics in practice. Section 3 presents the Value Sensitive Design framework used as the theoretical lens through which the research questions are addressed. Section 4 details the methodology used to prime the research analysis. Section 5 presents the empirical results that emerged from this analysis, the implications of which are critically discussed in Section 6. Finally, Section 7 concludes with a summary of the findings and future avenues of research.

2. BACKGROUND

2.1 Defining AI in the Public Sector

The ambiguity surrounding the definition of artificial intelligence continues to challenge researchers, practitioners, and policy-makers alike as there is still no universally accepted definition available for it (Grosz et al., 2016). 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. In particular, the European Commission, as of April 2021, presented a proposal for regulating AI. Because this paper inquires into the state of AI ethics in practice within the European context, it adopts the definition established by the European Commission in its proposal for regulating AI. 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).

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; Zuiderwijk et al., 2021). However, it generally lags behind the private sector in AI deployment (Mehr, 2017; Berryhill et al., 2019). As trends in big data and digitalization continue to rise, public service organizations are devoting resources to harness the power of data (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 that meet the General Data Protection (GDPR) requirements, encourages prolific AI development. Despite the recent developments in the field, empirical research on artificial intelligence in the public sector is limited (Sun & Medaglia, 2019). As a result, little is understood about the specific challenges of AI in the public sector, much less the ethical impact of AI (Aoki, 2020; Siau & Wang, 2020; Wirtz et al., 2020).

2.2 AI Ethics in Practice

Advances in AI and robotics have stimulated awareness and interest in 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 use, the ethics of AI becomes an important topic in terms of what society would look like in the future (Bartneck et al., 2021). A key issue in the field is defining to which ethical standards AI should adhere (Daly et al., 2019). In 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 embolden different stakeholders to carefully consider the ethics and associated morality of AI.

In terms of practicality, 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 a convergence of ethical principles appeared: transparency, justice, non-maleficence, responsibility, and privacy. However, critically important is the divergence that is observed, 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). They suggest that an alignment of ethical principles at the technology governance level can be achieved through standardization (2019). Yet, they raise the question as to whether these policy instruments have an impact on the practical implementation of AI or on the stakeholders upholding them. Particularly, do AI developers apply AI ethical guidelines in their practice? Hagendorff asserted that the adherence to principles outlined in ethical guidelines is poor in practice (2020). Furthermore, McNamara et al., in 2018 found that instructing software engineers to consider a code of ethics does not have a considerable, observed effect on their ethical decision-making. Thus, the onus of ethical decision-making does not solely rely on the individuals. Taking this further, Wirtz and Muller (2019) recommend setting up a public AI ethics committee to monitor the practical implementation of these standards.

On a macro-level, regulatory action as a stronger form of governance for AI has begun to appear as nations conceive their national artificial intelligence strategies. Smuha's article examines legislative tools available in the formation of AI regulation. However, the author states that regulators face the challenge of being subjected to self-governance elicited by ethical frameworks minus the lack of enforcement (Smuha, 2021). Notwithstanding, the European Parliament and Council have paved the way in terms of the first AI regulation. As of April 2021, the European Commission has released a proposal on AI regulation (EU Proposal AI Regulation, 2021). It also aims to harmonize the rules on AI in order to improve the AI ecosystem, and in general the economic markets.

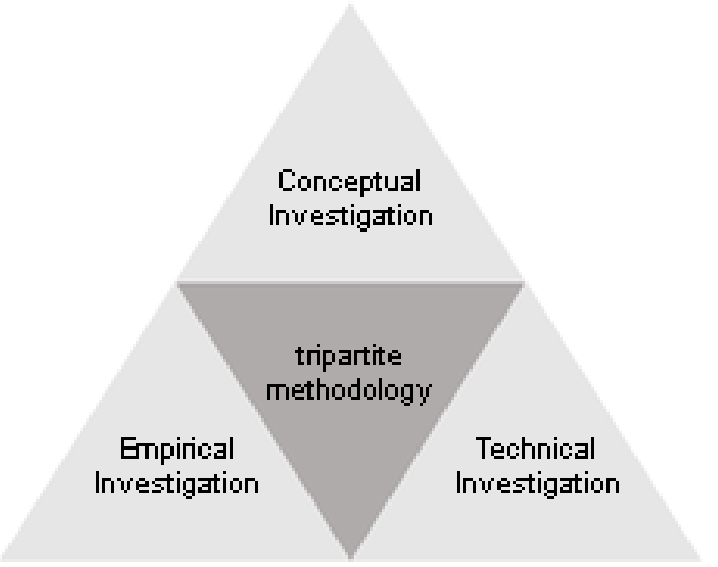
3. THEORETICAL FRAMEWORK: VALUE SENSITIVE DESIGN

The Value Sensitive Design (VSD) serves as the theoretical as well as methodological framework for this research. 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). VSD is suitable for 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). Embracing a sociotechnical approach, VSD draws from the human-computer interaction field. Furthermore, it has a characteristically tripartite methodology that combines conceptual, empirical, and technical investigations as shown in Figure 1.

The conceptual investigation is two-fold. On one hand, it explores the value source, implications, and trade-offs in a technology's design. On the other hand, it 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 the technology. Adjacent to this is the empirical investigation, which concerns the examination of the stakeholder's understanding, context and experiences relative to the technology and values. The empirical investigation can also inquire beyond the designer and into the organizational context of the AI and stakeholders. Completing the triad is the technical investigation, which inspects the technological properties, mechanisms, or features that may implicate the identified values and stakeholders. It focuses on the technology itself. (Friedman et al., 2002).

VSD has a wide range of beneficial features as a framework. 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. Furthermore, the methodology is iterative and integrative; it can be applied early in the design phase and throughout the process (Friedman et al., 2008, p. 85). Second, VSD emphasizes the need to identify both direct and indirect stakeholders, who, according to the authors, are often an afterthought in the overall design process (Friedman et al., 2008, p. 86). 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.

Figure 1.
Value sensitive design tripartite methodology



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, some differ relative to a particular cultural context and time period (2008, p. 86).

A critical weakness of the VSD is its 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, and virtue, to name a few, but does not commit to any one of them. In regards to VSD's claim of universality of values, Borning and Muller reject VSD's claims, calling its position on cultural relativism "problematic as well" (Borning & Muller, 2012, p. 1126). Instead, they suggest that VSD assume a pluralistic position that can then clarify "whether VSD is a method that can be applied to any set of values" (p. 1126).

Acknowledging the benefits and limitations of this approach, this research adapts the VSD method by complementing it with AI-centric ethical principles or values. In their paper AI4People, Floridi et al. (2018) synthesize five ethical principles that underpin the development and adoption of AI that serve the good of society as illustrated in Figure 2.

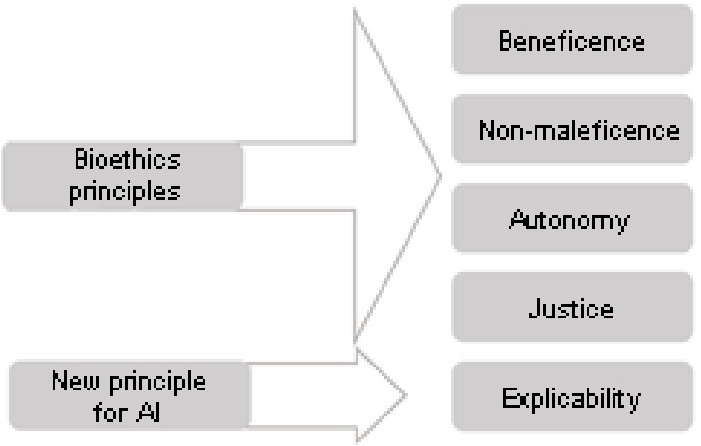
Beneficence: At its core, beneficence means promoting good in ethical terms (Jobin et al., 2019). Viewed as the common good, this principle concerns the promotion of well-being, preservation of human dignity, and sustaining the planet.

Non-Maleficence: Privacy, security, and safety are home to this principle. Privacy is closely related to the 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 pillar. 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. (2018) 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.

Justice: Under this principle are the concepts of equality, (non)-discrimination, accessibility, access and distribution, inclusion, and fairness among others (Jobin et al., 2019). More precisely, Floridi et al. (2018) indicate that justice can refer to a) using AI to correct past wrongs, b) ensuring

Figure 2.
Ethical framework for AI, comprised of five principles (Floridi et al., 2018, p. 700)



that the AI creates shared benefits, and c) preventing the introduction of new harms that exploit existing social structures.

Explicability: Accountability, transparency, comprehensibility, and interpretability are expressed under this principle in the sense of being able to understand what the AI does and why it is making the decisions it makes and holding such decisions or processes to account.

While the principles may not fully represent AI-implicated human values, nor do they claim any universality, the AI4People principles are a solid foundation to aid in the inquiry of AI ethics in practice. Thus, for this research, these two frameworks were selected to facilitate answering the research questions.

4. METHODOLOGY

4.1 Research Design

To discern how AI ethics is considered in practice by public service organizations, this qualitative study is guided by the Value Sensitive Design's (VSD) characteristically tripartite methodology: conceptual, empirical, and technical investigation. For this research, the tripartite methodology provides the pillars to support the translation of ethics into practice.

The conceptual investigation is divided into two components. First, participants and the involvement of parties are inquired about through the stakeholder analysis. This inquiry allows the values that play a role in the design and development of the AI to be extricated. 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. (2002) state that indirect stakeholders are left ignored in the design process. 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 development.

The empirical investigation explores the extent to which individual values are apprehended in the context of AI design and development and the extent to which these values are prioritized in design 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. Feedback from direct and indirect stakeholders about the AI is 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.

Selected as the country of focus, Estonia has over 70 identified use cases for AI in the public sector (Government of the Republic of Estonia, 2019). These 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 is 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 within the organization subject to feasibility testing or deployed for public use.

In addition, the use cases were limited to organizations that were available and agreed to this research on the condition of anonymity. The list of use cases is listed in table 1 in alphabetical order.

Qualitative data in the form of semi-structured interviews were collected from respondents from eight public service organizations that have developed an AI solution across the Estonian public administration. The respondents' roles varied from organization to organization, however, the commonality was their direct involvement in the design and development of the AI solution. Their roles are indicated in Table 2.

In total, data were collected from nine respondents representing the 8 public service organizations. The interviews were recorded and transcribed using an online transcription service. The transcriptions were independently reviewed for accuracy by the author. The anonymity of respondents was respected, thus identifiable characteristics were omitted to preserve confidentiality.

Table 1.
AI use cases by public service domain as of 2021

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

Table 2.
Interview respondents' roles

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

Coding was used to analyze the data collected. 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. 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, the analytical process was iterated to ensure the emergence of themes.

4.2 Research Limitations

The methodological approach of this research is subject to limitations. 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 contexts well beyond the borders of Europe which may be subjected to different measures, times, culture, and people.

Second, the unit of analysis is concentrated on the AI use case and the circumstances surrounding the design and development of the AI. Consequently, the perspectives offered on each of the use cases are significantly limited to these respondents' perspectives and may neither be reflective of the entirety of the AI project nor the organizational whole. Furthermore, most of the AI use cases were not completely developed or in full operational use at the time of research. Thus, a 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 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 emerged 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. Professional judgment by the author involved in the coding and analysis of values may therefore affect the interpretation of results.

5. RESULTS

5.1 Conceptual Investigation

Efficiency-Related Goals and Objectives: 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. Because the organizations had volumes of data that already existed, they decided to use their data in order to provide better services to their clients. The power of AI was also used to assess the efficiency of measures being implemented from policies and increasing the speed of delivering services, particularly in the public safety and emergency domain.

Immaturity of AI Solutions: The level of maturity of the AI solutions appeared consistently because, for a majority of the organizations, the maturity of their AI solutions was at the early stages of development or only implemented for use within the organization. The AI solutions were described as a "proof of concept", "a prototype", "trial phase", and "a pilot" project or phase. Some of the AI solutions were not used in production although the development of the prototype was completed, while others completed their first phase of trials. Others required additional work in the technical specifications of the solution, and some faced further data-related concerns.

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. The development of a prototype helped in determining not only feasibility but also establish a cost-benefit understanding. Featuring the most basic components required for its functionality, the prototype allowed the organization to experiment while managing costs. 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: The results showed that stakeholder involvement was limited to the development team and immediate users. Inadvertently, the possible impact of the AI solution's output on indirect stakeholders such as the organization, communities, or society at large was not considered in the design nor use because the solution was new, had too little data, or was in its early stages of development. Some teams used feedback forms from users to solicit areas for improvement of the tool. Of note also was the human supervision over the outputs delivered by the AI. Unanimously, there was an inherent understanding amongst the project teams that the human is ultimately responsible for any decision being acted upon as a result of the AI's output, yet it remained unclear whether this responsibility resided with the development team, the users of the AI, the head of the department, or collectively as an organization.

Transparency: 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 reached a consensus over whether the general public should be informed about the use of AI. All teams were in agreement that citizens should be informed about the AI's involvement in delivering a public service, regardless of whether it directly affected them. At the very least, the use of AI should be communicated to the public, whether through the terms of data processing agreements outlined in privacy policies.

5.2 Empirical Investigation

Data Governance and Usability: When asked about the challenges encountered when implementing ethics into practice, the major concerns shared by respondents were related to data governance and usability rather than the ethics or moral issues associated with it. Hard data – or data that came in the form of numeric values assigned to human traits such as success and achievement, motivation, intelligence, and violence factor – were difficult to translate into actual terms that would reflect nuances in reality. Consequently, such hard data would be used to train the AI solution. Limited, low-quality data also pervaded across organizations. Historically, the quality of the data was much lower than at present, and the AI solution required additional time processing the data to provide an output. Ensuring data compatibility also required a considerable amount of time and effort. Language-specific data to deliver a service in the Estonian language was not readily available, in comparison to English or Russian which had more speakers. Further restrictions also limited the use of an otherwise rich data lake that already existed within heavily-regulated organizations due to compliance with laws that limit how they can use and process such data.

Reliability and Maturity: Determining whether the output of the prototype is reliable was another pressing concern. Trust in the accuracy and dependability of the AI solution's output became a quality gate that restricted its use to test phases. Consequently, the ethical impact of the AI solution

could not be fully considered because of the maturity of the solution. This perspective revealed a correlation between the maturity of the AI solution and its ethical implications.

AI Skills and Competency: Building AI solutions within the confines of the public servants' own expertise became a challenge as they faced a steep learning curve. Thus, many organizations sought external assistance in the form of third-party AI vendors and AI advisors. However, even the tender process itself proved to be challenging as public servants and vendors refined requirements within the realm of feasible through multiple rounds. Public procurement specialists also 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. The guidance and expertise offered by the third-parties were bounded by project timelines and contracts. The responses inadvertently underlined the appetite for increasing technical AI competencies and skills.

5.3 Technical Investigation

Privacy: Where personal data was involved, special attention was given to privacy laws and how this would affect AI projects. Because the GDPR established privacy principles on minimizing data and limiting the purpose of data usage, these requirements were espoused by the organization through anonymization of training data or performing general-purpose analyses as opposed to citizen-centric analyses. Data protection impact assessments were also carried out on AI projects to demonstrate compliance. 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.

Security: The existence of personal data became a precondition for securing the AI solution itself and the processes supporting it. Because there was personal data obtained from data sources such as public registers, general security controls were applied through X-Road, which is Estonia's secured, centrally-managed distributed data-exchange layer. All data exchanged through X-Road was secured. Data not obtained through X-Road were housed in data centers that were protected by firewalls, access, and security controls. Other security measures were taken to secure the AI solution and its data such as controlling and restricting access. For example, a password and login combination were required to access the AI solution by those internal to the development team.

Automated Decision Making: Although some of the AI solutions had the ability to make decisions, most organizations purposely concluded any such automated decision-making with human review, oversight, and intervention. Meanwhile, some public services rendered could not be completely automated and thus sought AI as a decision-support tool that complemented human expertise. Respondents acknowledged that certain laws and guidelines advise against automatic decision-making by such tools. Respondents also added that due to technical limitations of their AI's capability and immaturity, automatic decision-making could not be achieved to the same extent a human would have done.

6. DISCUSSION

The VSD analysis reveals that a primary value driver for the design and development of AI in the Estonian public administration is the aim of achieving efficiency and effectiveness in public services. However, reaping the benefits of AI presents a challenge to governments as they tackle issues related to data governance, maturing of AI solutions, and AI skills, thus answering the first question of this research:

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

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 the computing performance of the AI, requiring higher computing resources. As a result, a considerable amount of time, costs, and effort is dwindled away by the preparation of data. Janssen et al. (2020) noted that this tedious task is given less consideration due to the time it takes. 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. For some organizations, the inability to use certain data for purposes outside of the initial terms can hamper the development of AI solutions. The lack of suitable data for training components of AI solutions adds a layer of complexity to the development process.

For the majority of the organizations in the study, the maturity of the AI applications seen was at the early developmental stages. AI solutions existed in the form of proofs-of-concept, 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.

The novelty of AI presents a steep learning curve for most organizations taking up AI initiatives. The lack of skills and technical competencies among public servants is clear as organizations sought guidance through engagements with third-party vendors specializing in AI technology implementation. Third-party vendors provide the technical expertise needed to design and develop AI solutions. Successful engagements can encourage future developments in organizations. However, procurement of these services proved to be a challenge. Because AI is new to most organizations, public servants are unfamiliar with navigating through the technical requirements and feasibility of building such solutions. Nonetheless, third-party vendors have a degree of influence over the outcomes of AI projects and ethical considerations throughout the design and development process.

6.1 Considerations for AI Ethical Principles

Initially, the results conveyed little to no consideration for the ethics of AI by public service organizations, owing to the immaturity of the AI solutions. However, the principles in action were activated to a certain degree, while some were more operationalized than others. The following subsections address the second research question:

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 showed that efficiency and effectiveness in order to improve the delivery of public services were the main values at play. Although not an ethical principle in and of itself, the intent was to deliver better quality services for the benefit of the citizens being served.

Non-Maleficence: This principle is manifested in tangible measures taken to ensure privacy, security, and safety. 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. Though not in service of AI ethical principles per se, these practices are a by-product of stringent regulations requiring such measures.

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 themselves. 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, but this awareness has the propensity to stem from data-related regulations, specifically GDPR's Article 22.

Justice: The stakeholders involved in the design and development of the AI solutions have been limited to direct stakeholders who often are small teams composed of people attentive to ensuring the

working 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, if at all. A lack of diversity in team composition and indirect stakeholder involvement may affect the way values are represented and consequently influence the design of the AI.

Explicability: The results of the interviews indicate that the black-box phenomenon is not prevalent. Public servants are able to explain how the AI solution arrived at its decision, citing that the same procedures could otherwise be performed using other tools albeit with more time and effort. In terms of transparency, the results conveyed that all the respondents 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 in use. 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.

The theme of the 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 the feasibility of the AI. But because the AI solutions are in such an early stage of development, the concern for risk and ethics is significantly diminished. Simply put, it is far too soon to describe its impact because the solutions are not fully developed or in use to cause harm yet. While there is some degree of awareness by public servants on the risks that are posed by AI, ethical guidelines or frameworks were minimally consulted. Taking all into consideration, these results shed light on 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 solutions that are aligned with the intent of improving public services for the benefit of public good. To some extent, AI ethical principles of beneficence, non-maleficence, justice, and explicability are indirectly considered and are somewhat practically demonstrated in a myriad of ways including: compliance with privacy regulations; the development of AI solutions with built-in security measures; a degree of awareness of the potential inaccuracy of the AI's output and how this may discriminate against certain groups or affect stakeholders; and openness for transparency when using AI to deliver public services to society. In this way, AI ethical principles are 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.

6.2 Implications and Recommendations

In light of the challenges that public service organizations face with AI applications as well as the limited practical implementation of ethical AI principles, the outcome of this research offers some guidance for further reflection. Designers of AI solutions should actively consider principles early in the design phase and throughout the development phase to reduce risks of unintentional harm. 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 potentially help address value conflicts and design AI solutions that are aligned with ethical values.

Governments should continue to develop a rich data ecosystem that enables sharing and exchange of high-quality data while maintaining security and integrity. Good data governance 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 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 in its infancy, while regulation of AI is on the horizon. Regulatory progress can provide guidance and direction in standardizing ethical principles and operationalizing them. Policymakers should examine the impact of proposed AI regulations on innovation and continue working with agility to calibrate legislation based on-the-ground input from all stakeholders and validate this with empirical data.

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 discussions on the ethics of AI, producing ethical frameworks and standards, and moving towards comprehensive regulation of the field. This research specifically takes on the topic of AI ethics by juxtaposing ethical concerns and the actual implementation of AI ethics in the public sector. More precisely, this research offers 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 the ethics of AI is being considered to a certain degree. Public service organizations indirectly translate ethical principles by way of addressing functional requirements by the organization and legal 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 are overshadowed by more pressing, practical issues related to the feasibility of AI solutions and data management.

Although a level of awareness on the risks posed by AI exists among public servants, their skills and competencies in the ethical development of AI can be further raised through training and various knowledge-sharing initiatives. While third-party AI vendors play a role in bridging this skills gap, 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 the sparse empirical scholarship on the ethics of AI. However, they are by no means sufficient to address the continuous debates on which stakeholder values and whose values are taken into consideration in the ethical development of AI. Therefore, suggested future areas of research on AI ethics in the public sector should examine citizens' perceptions of 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 cultural ideologies in different regions of the globe. 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.

ACKNOWLEDGMENT

This article would not have been possible without the generosity of the public servants and partners in the Estonian public administration who have willingly offered their time and knowledge in sharing their experiences on AI development.

CONFLICT OF INTEREST

The author of this publication declares there is no conflict of interest.

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). doi:10.1007/s11192-019-03222-9

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

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. doi:10.1007/978-3-030-51110-4_3

Berryhill, J., Heang, K. K., Clogher, R., & McBride, K. (2019). *Hello, World: Artificial intelligence and its use in the public sector*. 10.1787/19934351

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. doi:10.1145/2207676.2208560

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*. Advance online publication. doi:10.1016/j.giq.2020.101520

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. 10.2139/ssrn.3414805

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. doi: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. doi:10.1016/j.bushor.2019.11.004

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

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*(4), 689–707. Advance online publication. doi:10.1007/s11023-018-9482-5 PMID:30930541

Friedman, B., Kahn, P., & Borning, A. (2002). Value sensitive design: Theory and methods. *University of Washington Technical Report, 2*, 12.

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. doi:10.1002/9780470281819.ch4

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). doi: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. doi:10.1007/s11023-020-09517-8

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). doi: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. doi:10.1038/s42256-019-0088-2

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

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. doi:10.1145/3428502.3428513

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

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

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. doi:10.1007/s11948-021-00293-x PMID:33686527

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management*, *31*(2), 74–87. doi:10.4018/JDM.2020040105

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). doi:10.1016/j.giq.2018.09.008

Umbrello, S., & De Bellis, A. F. (2018). *A Value-Sensitive Design Approach to Intelligent Agents* (*SSRN* Scholarly Paper ID 3105597). Social Science Research Network. 10.1201/9781351251389-26

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

van Wynsberghe, A. (2013). Designing Robots for Care: Care Centered Value-Sensitive Design. *Science and Engineering Ethics*, *19*(2), 407–433. doi:10.1007/s11948-011-9343-6 PMID:22212357

Wirtz, B. W., & Müller, W. M. (2019). An integrated artificial intelligence framework for public management. *Public Management Review*, *21*(7), 1076–1100. doi: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. doi:10.1080/01900692.2020.1749851

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*(3). Advance online publication. doi:10.1016/j.giq.2021.101577

Charlene Hinton, CISA, CIPP/E, completed the Erasmus Mundus Master of Science in Public Sector Innovation and eGovernance (PIONEER) from KU Leuven with honors in 2021. As a senior consultant at an international firm, she holds multiple professional certifications in information security, information systems auditing and privacy. Her research interest includes the auditing AI systems, intersection of privacy, security, and AI, and the practical implementation of ethics in AI.