Alessandro Paciaroni

Trust Me, Trust Me Not

A Study on the Subjectivity of Estonian Civil Servants on the Use of Artificial Intelligence in the Public Sector

Master Thesis

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

Supervisor: Colin Van Noordt, Ph.D.
Tutor: Prof. Anu Masso.

Presented by: Alessandro Paciaroni

Date of Submission: 2022-08-09

Content

Tables

Abbreviations

ABV	Attention-Based View
AI	Artificial Intelligence
DOI	Diffusion of Innovation Theory
GPAI	Global Partnership on Artificial Intelligence
HLEG	High-Level Expert Committee
JRC	Joint Research Centre of the European Commission
ML	Machine Learning
OECD	Organization for Economic Cooperation and Development
TAM	Technology Acceptance Model

1 Introduction

Artificial Intelligence -i.e., AI- is a general-purpose technology; hence it allows for various uses that result in a context-dependent impact (Van Noordt and Misuraca 2020). This claim is consistent with the disruptive potential of intelligent technologies argued for by Bailey and Barley (2020). The scant use of Artificial Intelligence in the public sector adds to such difficulties (Wirtz et al. 2021). For this reason, research mainly focuses on the adoption of Artificial Intelligence by public sector organizations (Dwivedi et al. 2021; Sousa et al. 2019; Wirtz et al. 2021; Zuiderwijk et al. 2021). In this sense, the literature identifies factors influencing Artificial Intelligence adoption by public organizations on the environmental, institutional, organizational, and individual levels (Sun and Medaglia 2019; Van Noordt and Misuraca 2020).

Nonetheless, scholars tackled challenges to the study of phases following the adoption of Artificial Intelligence by public sector organizations in different ways (Campion et al. 2020; Sousa et al. 2019; Van Noordt and Misuraca 2020; Veale and Brass 2019). Overall, influencing factors in the post-adoption phases are identified at the same levels (Veale and Brass 2019). However, at the junction between adoption, implementation, and use of Artificial Intelligence, factors on the individual level are increasingly attracting the interest of scholars (Ahn and Chen 2021; Alshahrani et al. 2021; Horowitz and Kahn 2021; Sun and Medaglia 2019). This interest is consistent with the sociotechnical approach to studying technology (Guenduez et al. 2020; Meijer et al. 2021). In this context, research shows the importance of individual-level factors such as the perceived impact of technology on work tasks and the explainability of algorithms, as well as ethical concerns (Ahn and Chen 2021; Alshahrani et al. 2021; Horowitz and Kahn 2021). Notably, the research identifies trust as a pivotal factor (Alshahrani et al. 2021; Kumar et al. 2021; Thiebes et al. 2021). Accordingly, policymakers focus on developing trustworthy Artificial Intelligence to ensure its use by humans (Thiebes et al. 2021). However, trust depends on the human agent (Miller 2019; Zhu et al. 2021). While technology can be trustworthy, trust results from an estimate made by a human agent (Zhu et al. 2021). Specifically, trust is an estimate of trustworthiness (Zhu et al. 2021). Thus, there can be a mismatch between trustworthiness and trust deriving from preferences, experiences, and other subjective perceptions and characteristics of the trustor (Zhu et al. 2021). It derives that knowledge of the human component -i.e., the trustor (Zhu et al. 2021)- becomes pivotal to addressing trust in Artificial Intelligence -i.e., the trustee (Zhu et al. 2021). However, research on the matter either focuses on the perception of governmental employees on regulating the use of Artificial Intelligence in other fields (Horowitz and Kahn 2021); or on perceived challenges related to the adoption, assimilation, and implementation of Artificial Intelligence within the government

(Alshahrani et al. 2021; Kumar et al. 2021; Sun and Medaglia 2019). More tailored research on the perception of civil servants on adopting and supporting Artificial Intelligence within the work of public organizations is almost absent and with a generic focus (Ahn and Chen 2021). Overall, the importance of trust is simply pointed out concerning other factors (Kumar et al. 2021), and its connection with the perceptions of public sector employees on Artificial Intelligence is lacking (Ahn and Chen 2021).

Additionally, when the concept of trust comes into play, researchers and policymakers focus on the trustworthiness of technology, disregarding the human agent (Zhu et al. 2021). Therefore, there is a gap in extant knowledge on the perception of civil servants - i.e., the trustor (Zhu et al. 2021)- on the use of Artificial Intelligence within the government. This gap is relevant in that individual characteristics, and subjective perceptions of civil servants on the matter are pivotal to their trust in Artificial Intelligence (Zhu et al. 2021). To address this gap, this research aims to examine the subjectivity of civil servants regarding the use of Artificial Intelligence within the government with a focus on trust in Artificial Intelligence systems. In order to do so, the following research questions are formulated:

R.Q. 1 *What are the configurations of ideas about trust in Artificial Intelligence systems used in the public sector among Estonian civil servants?*

R.Q. 2 *How do Estonian civil servants configure ideas about trust in Artificial Intelligence with the pre-adoption phases of Artificial Intelligence in the public sector?*

R.Q. 3 *How do Estonian civil servants configure ideas about trust in Artificial Intelligence with the post-adoption phases of Artificial Intelligence in the public sector?*

Given the aim of this thesis and the insights on the use of Artificial Intelligence in the public sector, the Estonian government's use of Artificial Intelligence is chosen as the research setting. This choice derives from the wide use of Artificial Intelligence in the Estonian public sector. In particular, the use of Artificial Intelligence in the Estonian public sector includes use-cases among under-studied government functions (Sousa et al. 2019), namely Education, Defense, Health, Social Protection, Housing and Community Amenities, and Public Order and Safety (see https://www.kratid.ee/kasutuslood). This variety allows for the inclusion of different participants from different policy domains, which is considered an important factor in the context-dependency of using AI in the public sector (Vydra and Klievink 2019). In order to provide an answer to the research question, this research uses Q-Methodology. The foremost goal of Q-Methodology is the systematic study of subjectivity within a population and a given subject (Brown 1993). The main output of Q-Methodology shows how a group of participants interconnects -

i.e., configures- different elements of a subject matter (Brown 1993; Watts and Stenner 2005). This aims to contribute to the literature in that said configurations render how civil servants interconnect different elements of the debate surrounding the use of Artificial Intelligence in the public sector. Particularly so, in that this research emphasizes the aspect of trust in Artificial Intelligence. Additionally, practitioners involved in Artificial Intelligence projects in the Estonian public sector may find the results of this work valuable to prioritize and address specific facets of civil servants' perception of Artificial Intelligence or features of said technology.

This thesis is structured as follows. In the second section, the definition of Artificial Intelligence adopted in this thesis is introduced; the literature on the use of Artificial Intelligence in the public sector is examined, and the concept of trust in technology is presented. The third section explains the research approach and methodology and connects to the research philosophy. In the fourth section, the results of the data collection and analysis are presented. In the fifth section, the results are discussed in relation to the concepts and theories found in the literature; limitations are highlighted, and future research is suggested. In the sixth section, conclusions are drawn.

2 Literature Review

2.1 Defining Artificial Intelligence

AI is considered the new general-purpose technology (Van Noordt and Misuraca 2020). Therefore, its use spans many sectors and functions and fulfills many purposes (Van Noordt and Misuraca 2020; Wirtz et al. 2019). Scholars studying AI concerning the governmental function of governance differentiate it into three types: governance of AI, governance with AI, and governance by AI (Kuziemski and Misuraca 2020). Additionally, literature distinguishes the relationship between governments as a stakeholder and AI into two types: government as regulator and government as a user (Ballester 2021; Kuziemski and Misuraca 2020; Wirtz et al. 2019). The former refers to the so-called governance of AI. The latter is the focus of this study, which refers to governance with AI and by AI. Specifically, governance with AI is the object of this study. However, the next section will briefly show how governance with AI and by AI are profoundly related and intertwined, depending on the philosophical stance on technology.

Governance with AI relates to using AI in the public sector (Kuziemski and Misuraca 2020). Nonetheless, the situation in practice is more complicated. Governments have started using AI over the last few years and have done so slowly (Dwivedi et al. 2021; Sousa et al. 2019; Wirtz et al. 2019). Therefore, uses can differ quite a lot, and although research has shown the overall picture, there remains a lack of solid conceptualization and empirical validation of different uses. Generally speaking, research shows that AI is or can be used by governments in three main areas: policymaking, internal management, and provision of public services (Van Noordt and Misuraca 2022). More granular classification across the three areas is possible. For instance, agenda-setting and monitoring of implementation are different functions under the policymaking area, resource management and procurement are different uses under the internal management area, and provision of personalized public services and provision of new public services are different uses under the provision of public services area (Pencheva et al. 2020; Van Noordt and Misuraca 2022).

Following the surge of AI as a topic of public interest, a broader interest in defining its concept has emerged due to its application to business and government operations (Ballester 2021; Van Noordt 2022). As briefly mentioned and more thoroughly discussed in the context of other concepts relevant to this thesis in the following sections, the contamination of studies on the topic of AI by multiple disciplines does not facilitate common understanding even in core concepts, albeit providing heterogeneity and

richness (Ballester 2021; Van Noordt 2022). The substantial absence of a largely agreed definition of the concept of AI itself is a stark manifestation of this claim (Ballester 2021). Additionally, the broad interest in the topic leads to multiple conceptualizations of AI, each adopting a more or less different angle, showing the heterogeneity of contributions on the matter at hand. Although enriching the debate, this challenges research (Van Noordt 2022). Hence, this thesis needs to follow a definition that sets the scope of the matter of interest and delineates clear limits for generalization. Consequently, a conceptualization of AI is presented here to set the scope of this study. Given what has already been said on the heterogeneity of approaches to a definition of AI, the conceptualization used here consists of a combination of them. Nonetheless, only the ones relevant to this thesis are discussed here.

The European Union, through the work of the Joint Research Center on Artificial Intelligence -i.e., JRC AI Watch-worked on a definition of AI with an operational focus (Samoili et al. 2020). This effort could be ascribed to both an interest in knowledge creation and grounding the development of what became the first proposal of an AI Act (Samoili et al. 2020). Specifically, after the first version of the document, the EU proposed a legal definition of AI based on the risks posed by different systems in its proposal for an AI Act. Then the AI Watch published a review of its operational definition of AI motivating it with a need for a definition that can be used to detect what activities are considered AI, hence in scope for the AI Watch analyses (Samoili et al. 2020). For this discussion, it is important to focus on this rationale and the operational approach to defining AI. This approach is also consistent with the consideration made in the AI Watch report, according to which the approach used to classify AI depends on the study's objective (Samoili et al. 2020). Indeed, the work of the AI Watch results in a taxonomy that puts together core and transversal activities and topics of AI, and domains and subdomains of AI-related activities and topics (Samoili et al. 2020). The former categorizes activities related to the goals of specific AI systems -i.e., core activities- and activities or topics related to AI systems in general -i.e., transversal activities and topics (Samoili et al. 2020). The latter categorizes the domains and subdomains of said activities and topics related to AI (Samoili et al. 2020). Important in this thesis is the core domains of AI, identifying the core capabilities and goals of AI systems, and the transversal domain of integration and interaction, identifying the combination of two or more of the core domains with the environment (Samoili et al. 2020). The core domains proposed are reasoning, planning, learning, communication, and perception (Samoili et al. 2020). Among these, specifically relevant in this research, are learning, communication and perception. This choice is based on the considerations proposed in another research on the topic, whose explanation follows below. The approach used to classify AI ultimately responds to the study's goal (Samoili et al. 2020). It is to say that the elements considered

for the classification depend on the dimension deemed relevant for defining the topic at hand. Therefore, the heterogeneous interests in AI result in multiple definitions that look at AI from different lenses (Van Noordt 2022). Additionally, and as a result of the lack of a large agreement on the definition of AI on the academic side, it appears necessary to consider the perspective of key stakeholders (Van Noordt 2022). In this sense, public organizations are considered the most important stakeholder due to the focus on the use of AI by the government (Van Noordt 2022). Noticeably, the approach taken by the AI Watch provides insights on this. Nonetheless, the focus of said definition is to determine which activities are associated with AI, thereby being worth analysis (Samoili et al. 2020). In contrast, this thesis focuses on a narrower aspect of the field. Namely, the use of AI by public sector organizations and the subjective perceptions of individual civil servants on such technology. However, research defining AI from this specific angle is scarce (Sousa et al. 2019; Wirtz et al. 2021; Zuiderwijk et al. 2021). Nonetheless, some of the extant approaches in the literature appear to be particularly suitable for the objective of this thesis (Van Noordt 2022). Additionally, these are consistent with some of the dimensions of AI highlighted by the AI Watch (Samoili et al. 2020) and with the prevalent perceptions of public servants on AI (Van Noordt 2022). Conceptualizations of AI in the literature range from futuristic ideas on superintelligent non-human agents to the science behind the development of AI (Van Noordt 2022). In between, among others, are perspectives that conceptualize AI with a technology-intensive, capabilities-centered, and applications-based focus (Van Noordt 2022). The three latter conceptualizations are relevant to this thesis. Indeed, conceptualizing AI based on the techniques used, albeit sometimes imprecise, is of scientific relevance for this research as not every technology commonly associated with AI has unique elements compared to other technologies (Van Noordt 2022). For instance, Bayesian methods may be associated with AI despite being probabilistic statistical techniques, albeit complex (Van Noordt 2022). Case-based reasoning is also associated with AI; however, this technique automates reasoning but relies on the rules set by the developer; hence it does not fall far from the basics of every software operating through an IF-THEN functionality (Samoili et al. 2020). On the contrary, Machine Learning is arguably the most famous AI technique, renowned for allowing an AI solution to learn by itself based on a given dataset (Samoili et al. 2020). Worth of mention is the fact that also within Machine Learning there are different levels of autonomy of the learning process: depending on whether this is supervised or not, the developer will have a role in what the AI solution takes into account during the learning process (Samoili et al. 2020). In this research the focus lies on this technique as the autonomy of the AI solutions arguably creates a situation of uniqueness in comparison to other, albeit advanced, technologies already used in the public sector for a long time. Nonetheless, the main limitation to this approach is the dynamicity of this element of AI

(Samoili et al. 2020). Indeed, in the past symbolic reasoning was considered to be AI whereas now is mostly associated with rule-based algorithms and may not fit the current definition of AI (Samoili et al. 2020). This is why considering additional conceptualizations is deemed necessary. Particularly relevant appears to be the capability held by an AI solution (Van Noordt 2022). This is consistent with the perspective of Belgian civil servants (Van Noordt 2022). In this case, perceiving, learning, and making decisions are of interest to this thesis as they are commonly associated with humans, albeit being shown by a non-human agent (Van Noordt 2022). It is to say that this is another element of the uniqueness of AI and it is deemed of pivotal importance in shaping the perception of human agents. Lastly, these capabilities are usually associated with specific applications (Van Noordt 2022). Logically, human agents have in AI applications their first point of reference for what AI consists of, as shown by the views expressed by Belgian public servants (Van Noordt 2022). In this case, recommendation softwares are of particular interests as their function consists of producing outputs that the human agent is called to act upon. It is important to notice that this is not a unique feature per se. For instance, traditional ERP systems might provide recommendations for action to the user (Samoili et al. 2020). However, it is the combination of the conceptualizations considered here and their specific elements that make the source of the recommendations unique (Van Noordt 2022). This last point is also the key rationale for this thesis' chosen approach to conceptualizing AI.

Provided the account of the challenges and objective dependence of the definition of Artificial Intelligence (Ballester 2021; Samoili et al. 2020; Van Noordt 2022), Artificial Intelligence shall be defined for this thesis. This task follows a combination of conceptualizations on multiple dimensions. Specifically, this is drawn from Van Noordt (2022) due to the consideration of the perspectives of civil servants made by the author, consistently with Samoili et al. (2020). Therefore, Artificial Intelligence is defined to scope this thesis as follows. Artificial Intelligence is defined from the perspective of its technologies, capabilities, and applications (Van Noordt 2022). As seen in Samoili et al. (2020), AI systems are likely to combine more techniques and capabilities, thereby making it unfeasible limiting the scope to AI systems relying solely on a specific technique or expressing only one capability. Artificial Intelligence systems considered in this thesis must encompass Machine Learning techniques in the technological dimension. In the dimension of capabilities, AI systems considered in this thesis must include a combination of perceiving, learning, and decision-making capabilities (Samoili et al. 2020; Van Noordt 2022). On the dimension of specific AI applications, this thesis considers recommendation softwares as the recommendations based on the abovementioned capabilities and techniques are unique to AI (Van Noordt 2022). To summarize, Artificial Intelligence systems considered in this thesis are recommendation

softwares expressing a combination of perceiving, learning, and decision-making capabilities based on Machine Learning.

2.2 Theories and Concepts on the Use of Artificial Intelligence in the Public Sector

A crucial angle in the research of AI use in government focuses on what the use of AI implies for governments. Here three units of analysis can be found in the literature: the individual -i.e., impact on public servants' work- the organization -i.e., impact on organizational structure and processes- and the institution -i.e., the bureaucratic form and form of governance. Literature in this area can be distinguished based on the rationale that it is based on. It is to say that the philosophical standpoint on technology, hence the theories used by different scholars, determines two main viewpoints on what the introduction of AI in the government implies for both public servants' work and organizational structure and processes. The two philosophical standpoints are constructivism and determinism. To put it simply, the former claims that technology is socially constructed; hence its use depends on the context and social norms and values; the latter affirms that technology itself determines its application; hence its use depends on the features of the technology and organizations and individuals are impacted accordingly. Categorizing contributions according to a more granular distinction of philosophical viewpoints is out of the scope of this study. Nonetheless, it is essential to touch on the matter as its discussion provides insights into available knowledge on AI use in the public sector.

Regarding approaches leaning towards determinism, technological adoption by the government and the penetration and diffusion within and across public sector organizations are deemed to have a deep impact on both the nature and modalities of work of individual civil servants (Bovens and Zouridis 2002; Young et al. 2019) and on the public sector organization as a whole (Lorenz et al. 2021; Pencheva et al. 2020). AI and algorithms bring unique affordances to decision-making tasks (Young et al. 2019). This impact affects information sources and processing, coordinating mechanisms, and authority, thereby shaping a new ideal-type of bureaucracy (Lorenz et al. 2021). Changes in the work of civil servants and governmental organizations started already with the first introduction of ICTs (Bovens and Zouridis 2002; Lorenz et al. 2021). Indeed, using ICTs in the public sector shifts from the so-called street-level bureaucrats to screen-level bureaucrats (Bovens and Zouridis 2002). The former are employees such as social workers implementing policies on a case-by-case basis; the latter are employees with a similar role but now handling the cases through their screens, hence having to follow crystallized rules and procedures (Bovens and Zouridis 2002). Subsequently, the

extension of ICT into large information systems produces a shift toward system-level bureaucrats -i.e., employees such as system analysts and engineers shaping the policy implementation by designing and maintaining large information systems – (Bovens and Zouridis 2002). This trend is found to influence discretion in decision-making (Bovens and Zouridis 2002; Bullock et al. 2020; Young et al. 2019). Discretion is influenced in that the use of technology affects authority within public sector organizations (Lorenz et al. 2021). It is to say that the level of organization where authority is exerted shifts; specifically, it does so upwards in the hierarchy (Bovens and Zouridis 2002; Lorenz et al. 2021). Authors identify evidence of these changes in a set of characteristics of the public sector. Bovens and Zouridis (2002) highlight the role and functions of ICT, the discretion of street-level bureaucrats on individual cases, and the roles of key professionals as loci of change, among other factors. The authors identify that the role of ICT becomes decisive in the so-called system-level bureaucracy, as the functions of ICT range from execution and control to communication (Bovens and Zouridis 2002). Hence, the systems designer becomes the authoritative role, and discretion on a case-by-case basis is absent (Bovens and Zouridis 2002; Lorenz et al. 2021). With regards to the characterization of this fashion of bureaucracy, it is worth noticing that despite authority shifts upward, public-sector employees still hold the steer of the organization. Indeed, drawing on Zuurmond, Lorenz et al. (2021, p. 75) claim that "he who controls the information systems, controls the organization." However, a deeper impact on the whole policy cycle is identified with the introduction of new technologies such as big data and AI (Pencheva et al. 2020). The difference is due to the affordances associated with AI (Young et al. 2019). For instance, information is the basis of decision-making for every type of government (Lorenz et al. 2021). However, scholars distinguish between the use of information by governments after the introduction of ICT and after the introduction of algorithms, which calls for a new type of discretion (Young et al. 2019). Within decision-making, artificial discretion is different from human discretion in that the former increases the scalability and quality of decision-making tasks and decreases their cost (Young et al. 2019). Indeed, ICT allows for efficient use of information across governmental organizations, whereas the boundaries of the information basis shift outside governmental organizations with the introduction of big data and AI (Lorenz et al. 2021). This affordance is deemed to have a profound effect on the policy cycle; for example, it allows for more accuracy, legitimacy, and collaboration across governmental boundaries in the agenda-setting and policy formulation phases (Pencheva et al. 2020). The same applies to how information is processed in public sector organizations in that advanced technologies such as AI can be used for non-routine tasks, whereas ICTs are built on rule-based algorithms that enable more efficient execution of routine tasks (Lorenz et al. 2021). This change in information processing would produce a shift in phases such as the policy implementation, where even

non-routinized decision-making can be executed by advanced technologies (Pencheva et al. 2020). Furthermore, the prevalent coordinating mechanism when advanced technologies are used lies in their automated advice whereas ICT-based governments rely on information systems controlled by said system-level bureaucrats (Bovens and Zouridis 2002; Lorenz et al. 2021). Eventually, these affordances influence where the locus of authority is in the government, which Lorenz et al. (2021) individuate in the highest level of the organization as the authority is exercised through the algorithm itself. Elaborating further on the analysis of the impact of advanced technologies on governmental organization a scenario similar to the so-called governance by AI would be possible, where algorithms govern society, and the democracy transforms into a technocracy (Newman et al. 2022).

Following a similar approach, the same and other scholars claim that the extent of the impact of the use of AI in governments is contented (Newman et al. 2022) and deemed likely to be context-dependent (Vydra and Klievink 2019). The nature of tasks at the individual, organizational, and institutional levels varies, making tasks more or less likely to be impacted by the use of AI (Bullock et al. 2020). Influencing factors on the micro- meso- and macro-levels other than the nature of tasks are found in the literature. They consist of behavioral mechanisms, human-computer interaction design, policy domain, and bureaucratic form and tradition (Ahn and Chen 2021; Criado et al. 2020; Meijer et al. 2021; Peeters 2020; Vydra and Klievink 2019). The introduction of new technologies has been moving the authority towards system designers and reduced human discretion, as argued by Bovens and Zouridis (2002), for highly automatized tasks in the general service provision (Sousa et al. 2019). However, it is also true that, for instance, policy formulation demand high discretion, and the impact of new technologies is low to date (Vydra and Klievink 2019; Young et al. 2019). Bullock et al. (2020) analyze the impact of AI concerning both system-level bureaucracy and artificial discretion. The resulting classification sees tasks with a low level of discretion required as a system-level form of bureaucracy and highly likely to be characterized by artificial discretion (Bullock et al. 2020). In contrast, tasks requiring high discretion are not likely to be deeply affected by this shift in discretion (Bullock et al. 2020). The classification of the bureaucratic form is a function of the general nature of tasks in the organization (Bullock et al. 2020). This in turn, is a function of the policy domain (Bullock et al. 2020; Vydra and Klievink 2019).As a consequence, the impact of AI is a function of the bureaucratic form (Bullock et al. 2020). Meijer et al.'s (2021) findings support and extend this claim through an in-depth analysis of the use of a predicting policing algorithm in two different organizations. The different use of the same algorithm by the German police in Berlin and the Dutch one in Amsterdam shows how social norms and interpretation shape the use of technology (Meijer et al. 2021). These findings confirm the influence of the bureaucratic form on the

use of AI, hence its impact on work and extends the causal relationship to the social norms, values, and beliefs upon which the bureaucracy is built (Meijer et al. 2021).

Considering social norms and the interpretation that civil servants give of AI moves the focus of explanatory factors of variance of use and impact of technology to the very social component of the relationship between humans and technology. Meijer et al.'s (2021) study proves that this is relevant across (administrative) cultures, whereas Vydra and Klievink (2019) extend its validity to cross-policy domains variance. The issues at hand in specific policy domains imply different rationales: solving complex social issues, such as criminal recidivism, is associated with finding the root cause of the problem, whereas, in other contexts, correlation based on ML data analysis is more readily accepted by decision-makers (Vydra and Klievink 2019). Furthermore, psychological factors, values, beliefs, and expectations influence variance at the individual level (Peeters 2020).

In conclusion, it is not the intention of this thesis to solve this philosophical debate; hence the discussion of this matter shall not continue further. Moreover, recent studies provide an overarching overview of the factors and variables underpinning different results of the use of intelligent technologies (Bailey and Barley 2020). These range from the ideology of the designers to the characteristics of social institutions where technology is applied and go from the ideation phase to the long-term effects on society (Bailey and Barley 2020). For this reason, the literature discussion should continue with an overview of extant knowledge on the factors and variables pertaining to the phases relevant to this study. Therefore, the following two subsections will address relevant factors on the environmental, organizational, and individual levels of analysis on the phases of adoption and post-adoption, respectively.

2.2.1 Pre-Adoption Phases of Artificial Intelligence in the Public Sector

Relatively extensive knowledge on AI adoption by governments is available, in contrast to knowledge of the post-adoption phases (Sousa et al. 2019; Wirtz et al. 2021). Scholars point to the fact that AI in the public sector is a relatively new phenomenon; hence research is mainly focused on the early phases of the technology's use in the organization (Sousa et al. 2019). In addition to the relevance of the phases leading to the adoption of AI by governments, it seems that another reason why research focuses on these stages is that the hype surrounding this new technology elicits theoretical reflection and development of AI-specific frameworks, often resulting in the adaptation of existing frameworks (Wirtz et al. 2021). The conceptual focus may be due to the uncertainty surrounding general-purpose technology, especially in the first phases of its emergence (Wirtz et al. 2021). Research with this focus is mainly based on (multiple) case studies and expert interviews (Wirtz et al. 2021).

AI is deemed a force for change in the public sector (Bullock et al. 2020; Lorenz et al. 2021; Van Noordt and Misuraca 2020; Veale and Brass 2019). However, the potential change is bound to antecedent factors that may drive or hamper innovation (Van Noordt and Misuraca 2020). Scholars foresee the possibility of disruptive changes along the whole policy cycle resulting from using new technologies, such as big data and AI (Pencheva et al. 2020). Therefore, the range of antecedents spans from individual to environmental factors (Van Noordt and Misuraca 2020; Veale and Brass 2019). The environmental context influences innovation in the public sector (Van Noordt and Misuraca 2020). This influence is exerted by both public and private actors (Van Noordt and Misuraca 2020; Wang et al. 2020). On the one hand, public sector organizations exert pressure on each other when similar organizations strive to achieve a competitive advantage to reach higher performance levels, especially in local governments (Wang et al. 2020). On the other hand, innovation diffusion occurs through isomorphism -i.e., a tendency to alignment- (Van Noordt and Misuraca 2020). Intergovernmental cooperation and agreements on AI principles and good practices manifest such tendencies (see: AI HLEG, OECD AI, GPAI). Interorganizational cooperation is also relevant on a local level in that it is associated with a better possibility of harnessing the benefit of AI (Campion et al. 2020). Therefore, public and private networks play a role as an environmental antecedent of AI adoption (Van Noordt and Misuraca 2020). The environmental push for inter-organizational cooperation is tightly coupled with institutional challenges. Van Noordt and Misuraca (2020) find that inter-organizational data sharing within the public sector and between the public and private sectors is an important environmental antecedent. However, Campion et al. (Campion et al. 2020) highlight how data sharing is a concern for governmental organizations. This is particularly true because of the sensitiveness of data collected and used by governments (Campion et al. 2020; Pencheva et al. 2020). According to Veale and Brass (2019), the concerns on data sharing are one of the key elements underpinning the development of strengthened coordination mechanisms at the institutional level, such as the UK Centre for Data Ethics and Innovation. In this context, laws, regulations, and frameworks play an important role at the institutional level. For instance, in the study of an algorithm for internal auditing in the Spanish public sector, Criado et al. (2020) found that the clear legal definition of functions and limits of the algorithm has a positive effect on its use by the civil servants and their discretion. Additionally, new capacities and capabilities are associated with the effective adoption and deployment of new technologies like AI and ML (Mikalef et al. 2021; Veale and Brass 2019). For instance, technology strategies are being developed to frame and steer the adoption and implementation of such technologies (Veale and Brass 2019). This antecedent undeniably poses demands on the organizational level in that policies are implemented at this level (Veale and Brass 2019). Indeed, Wang et al. (2020)

find that another factor influencing AI adoption by local governments is vertical pressure -i.e., internal pressure from higher levels in the hierarchy. Organizational antecedents consist of both structural and cultural factors (Van Noordt and Misuraca 2020). Capacity and capability building, as well as changing processes and working routines at the organizational level have an impact on AI adoption (Campion et al. 2020; Pencheva et al. 2020). Van Noordt and Misuraca (2020) argue that data sharing be considered an antecedent in itself, given its great impact on AI adoption. Nonetheless, it seems that it is a crucial factor influencing organizational antecedents. Campion et al. (2020) observe that concerns regarding data privacy and security, and a lack of understanding of data needs are among the key challenges to interorganizational cooperation on AI adoption projects. In particular, the authors highlight the importance of interorganizational trust as a factor influencing the effective data sharing required by such projects. Organizational resistance to data sharing is highlighted also by key stakeholders in public healthcare (Sun and Medaglia 2019). On the one hand, these elements pertain to the sphere of organizational culture (Campion et al. 2020). On the other hand, they require organizations to review their processes, infrastructure, and routines (Pencheva et al. 2020), thereby directly impacting its structure. For instance, Criado et al. (2020) find that co-development of AI systems positively influences the adoption and use of such technology by first-line bureaucrats. The authors also identify cooperation between management units as a tool to facilitate its adoption. It follows that management support is another relevant antecedent, together with organizational resources (Sun and Medaglia 2019; Van Noordt and Misuraca 2020). As seen the social component of the socio-technical system is an important factor to explain different results of AI adoption, use and impact on the public sector across levels of analysis (Meijer et al. 2021; Peeters 2020; Pencheva et al. 2020; Van Noordt and Misuraca 2020; Veale and Brass 2019). Harnessing the potential of advanced technologies such as big data, ML and AI is dependent on framing them into policies (Veale and Brass 2019). On this basis such technologies may be adopted with different goals (Veale and Brass 2019). In this context the perceptions and expectations that decision makers have on AI are a pivotal factor: indeed, adoption and effective use of big data and ML is found to be bounded to a shift in the mindset of policy makers and public managers (Pencheva et al. 2020; Pi 2021). Framing AI systems in some ways rather than others is likely to influence the result of the adoption. Criado et al. (2020) show that in adopting an internal auditing algorithm, its definition as a decision support system instead of a decision-making system is likely to have helped the acceptance of civil servants, hence its adoption. This is also related to indirect environmental factors in that conceiving AI as a well-suited solution to external triggers facilitates its adoption (Van Noordt and Misuraca 2020). On the contrary, Sun and Medaglia (2019) identify unrealistic expectations on and misunderstanding of AI as

challenges to its adoption. Nonetheless, this is not a direct relationship in that scholars identify moral and ethical considerations and trust to influence the adoption of AI in the public sector (Kumar et al. 2021; Pencheva et al. 2020; Pi 2021; Sun and Medaglia 2019). In particular, Kumar et al. (2021) identify a lack of trust in technology as a foundational barrier to AI adoption in post-distribution systems in the Indian public sector. It is to say that, according to the model used by the authors, a lack of trust in technology is driving other barriers.

2.2.2 Post-Adoption Phases of Artificial Intelligence in the Public Sector

Knowledge of the post-adoption phases of AI in the public sector is largely based on the use of theories of technology and innovation diffusion (Alshahrani et al. 2021; Mikalef et al. 2021). As the theories underpinning the pre-adoption phases, theories of technology and innovation diffusion identify environmental, organizational, and individual factors impacting AI use and appropriation by public sector organizations. Nonetheless, research on the post-adoption phases of AI in the public sector is little compared to previous phases (Dwivedi et al. 2021). On the one hand, the reason is that AI use in the public sector is scant and has increased only recently, thereby making researching these phases less feasible (Dwivedi et al. 2021). On the other hand, building knowledge on these phases must consider intervening factors prior to the adoption phase and complementary factors in the post-adoption phases (Bailey and Barley 2020; Van Noordt and Misuraca 2020). Moreover, as stakeholders change along the technology life cycle, factors influencing pre-adoption and post-adoption phases take different fashions or change altogether (Zhang et al. 2021). Additionally, further distinctions between factors influencing different post-adoption phases are identified (Campion et al. 2020).

Environmental factors in the pre-adoption phases are identified in the role of networks (Van Noordt and Misuraca 2020). However, in the post-adoption phases, the role of pressure from citizens gains importance (Mikalef et al. 2021). Additionally, the role played by the external service environment increases in importance as it does not relate only to the decision to adopt AI anymore but also to the quality and effectiveness of AI-based service provision (Zhang et al. 2021). The focus on the quality of AI-based public services is consistent with the relevance of public value creation by the public sector - measured by performance, openness, and inclusion- which is highlighted and used by scholars to frame the research on the impact of AI in public sector organizations (Van Noordt and Misuraca 2020). No major differences between pre- and post-adoption of AI can be found at the institutional level. This similarity is consistent with the fact that the broad frame for policy formulation is set at this level, whereas the implementation and subsequent phases are more influenced by organizational and individual factors (Veale

and Brass 2019). Nonetheless, it shall be re-stated here that supportive policies and regulations are positively related with successful implementation and use of AI (Campion et al. 2020; Mikalef et al. 2021; Zhang et al. 2021). For instance, well-suited data sharing agreements and feedback channels are identified as influencing the implementation phase as well (Campion et al. 2020; Zhang et al. 2021). Factors on the organizational level acquire slightly different fashions as the post-adoption phases entail the implementation of AI in the organization business processes and daily work routines, whose success is associated to the development of organizational capabilities (Mikalef et al. 2021). Nonetheless, scant research is available on the matter. For instance, Campion et al. (2020) analyses the challenges of the implementation phase to fill this gap and find that lack of awareness on AI potential is a prominent lack in the organizational skills in this phase, whereas the lack of leadership skills is more relevant in the adoption phase. Additionally, in terms of organizational culture distrust is identified as a challenge for successful implementation of AI (Campion et al. 2020). This is consistent with the relevance of collaborative development of AI systems and alignment of interorganizational needs highlighted as positively influencing adoption and implementation phases (Campion et al. 2020; Criado et al. 2020). Public managers generally have autonomy in the implementation of AI in public sector organizations, and civil servants have a direct interaction with such technologies in their daily work (Veale and Brass 2019). Therefore, the influence of individual factors acquires greater importance in the post-adoption phases. For instance, in addition to organizational culture mediating decisions on adoption of AI, the attitudes of individual public servants towards the specific AI system are found to have a greater impact in post-adoption phases as this influences the daily use of such technology (Ahn and Chen 2021). The same holds for the perceptions that public managers have of AI and algorithms: Veale and Brass (2019), for instance, identify a trade-off between accountability and quality assurance that public managers are called to solve in the implementation and use of these technologies. A theoretically informed account of the relationship between organizational and individual factors is provided by Alsharani et al. (2021). The authors highlight the shortcomings of adoption models, namely an overweight of the autonomy of individuals in the adoption of innovation that does not sufficiently account for the role of authority and contingency in the adoption of innovation (Alsharani et al. 2021). To solve this, they adopt an Attention Based View that posits that what decision-makers do depends on what they focus on, which depends on the context where decision-makers find themselves, which depends on the structural and cultural features of the organization (Ahn and Chen 2021). Expectations and other psychological attitudes are highly relevant in the post-adoption phases (Ahn and Chen 2021; Guenduez et al. 2020; Meijer et al. 2021; Zhang et al. 2021). Ahn and Chen (2021) provide and apply a conceptual framework that integrates a technology acceptance model

-i.e., TAM- and a framework of innovation diffusion -i.e., DOI- to fill the knowledge gap on attitudes of civil servants toward AI. The authors find that both perceptions of the usefulness of AI in their daily work and on the societal impact of AI, as well as moral and ethical considerations, influence the willingness of civil servants to support the use of AI in the organization (Ahn and Chen 2021). Perhaps the importance of expectations in creating perception is best explained by the account that the authors provide in their assessment of the influence of knowledge on the willingness to use and support AI. The authors suggest that a pattern of agreement and strong agreement in the categories of the survey questions may be explained by respondents' relatively scarce knowledge of AI technology -namely, civil servants using AI in their work- (Ahn and Chen 2021). Importantly, the authors argue that although they tried to control this factor by measuring the level of knowledge, this was also a perceptual measure (Ahn and Chen 2021). Thus, it is argued here that there is a case in point to assume that since even the measurement of knowledge of AI is based on perceptions, there is a relation between the expectations of users on what the technology is used for and how it works and that this influences the perception of being knowledgeable about it.

Additionally, it should be noted that research on the impact of AI in public sector organizations fits under the research strand that focuses on the post-adoption phases. For instance, scholars analyzed algorithms' impact on the public sector regarding authority, information use and processing, coordinating mechanisms, and artificial discretion (Bullock 2020; Lorenz et al. 2021). However, it is crucial to note that this literature addresses the potential impact of the use of AI and algorithms in the public sector but provides insufficient knowledge of their potential enablers or bottlenecks, thereby bounding a final assessment to future research (Newmann et al. 2022). Therefore, guidelines for successful implementation and use of AI in the public sector are primarily based on the first strand of research presented in this section -namely, research on pre-adoption and adoption phases- (Thiebes et al. 2021). These guidelines consist of practice-oriented recommendations on how to address perceived risks related to AI and practically develop trustworthy Artificial Intelligence (Thiebes et al. 2021). It is argued here that, albeit necessary, this is not enough for effective implementation and use of AI in that this research underweights the role played by the subjectivity of expectations and attitudes of civil servants in the post-adoption phases. The following section introduces and explains these guidelines, namely the Ethics Guidelines for Trustworthy AI -henceforth, Trustworthy AI- developed by the AI High-Level Expert Group of the European Commission -henceforth, AI HLEG-, it reviews theories and concepts relevant to the guidelines and this thesis, and it highlights the knowledge gaps that this research aims to fill.

2.3 Trust as a Success Factor for the Use of Artificial Intelligence in the Public Sector

As presented in the previous sections, the exponential improvements in machine learning, the bread and butter of AI, have prompted exploratory research (Wirtz et al. 2021). Among the extant research, an increasing number of scholars focus on this technology's (perceived) potential, both in terms of benefits and threats (Thiebes et al. 2021). The rationale of this research is expressed by the AI HLEG, as reported by Thiebes et al. (2021), as the idea that humans will ever be able to use AI to its full potential only if there is trust in the technology. Thiebes et al. (2021) also argue that the background of disciplines informing this research is heterogeneous, suggesting its richness albeit recognizing its fuzziness. Expanding on this, the authors pose that literature on trustworthy AI focuses on technical and social factors, thereby exhaustively accounting for elements of trust (Thiebes et al. 2021). Trustworthy AI supports their claims, a set of guidelines developed by AI HLEG to provide guidance on how to develop a Trustworthy AI system. It is to say that the structure of the guidelines effectively includes the dimensions of trust conceptualized by scientific literature. For exposition purposes, Trustworthy AI of AI HLEG will be briefly presented now, whereas the following sections will address the literature on trust, highlighting the limitations of the guidelines concerning the academic knowledge on trust.

Trustworthy AI of the AI HLEG is constructed as follows: lawful, ethical, and robust AI are the pillars of trustworthy AI. However, lawfulness is not explicitly referred to in the guidelines. Drawing upon ethical AI are four ethical principles: respect for human autonomy, prevention of harm, fairness, and explicability. Eventually, these principles underlie seven essential requirements: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity non-discrimination and fairness, societal and environmental wellbeing, and accountability. Aside from assessing that both social and technical elements are present in the requirements, Thiebes et al. (2021) argue for the exhaustiveness of the AI HLEG Trustworthy AI guidelines because they identify the presence of the ethical principles proposed by Floridi, arguably an authoritative voice in the field, in the guidelines. Notwithstanding the relevance of the guidelines for practice and the unavoidable bias introduced by ethical considerations in the highly politicized field of AI, it can be claimed that the views and assessments presented here have significant limitations. Notably, the critique moved here is far from deriving from other ethical predispositions, and it is grounded on empirical research and a tweak in the conceptual underpinnings of the literature on trustworthy AI.

The AI HLEG guidelines are relevant and fit as far as trustworthiness is in scope. However, scholars distinguish between trustworthiness and trust (Robbins 2016; Zhu et al. 2021). What is highlighted here is conceptual confusion between trustworthiness and trust. The distinction is made clear by Robbins (2016) in a discussion on the unclarity of the concept of trust that focuses on trust between individuals. Here the author identifies trust as a cognitive process of the truster that believes in the trustworthiness of the trustee (Robbins 2016). The relation between the concept is a function: trust is a function of the trustworthiness of the trustee (Robbins 2016). Ultimately, trustworthiness is the capability and commitment of a trustee to behave accordingly to the expectations of a truster (Robbins 2016). A similar clarification appears to be necessary for the field of AI as well, according to Zhu et al. (2021). Expressly, the authors point to this misinterpretation of the concepts of trust and trustworthiness as one of the main reasons for concern in the research on the operationalization of AI ethical principles (Zhu et al. 2021). Additionally, relevant for this thesis, *mutatis mutandis*, they extend the validity of the conceptualization proposed by Robbins (2016) to the field of AI. It is to say that Zhu et al. (2021) distinguish the trustworthiness of an AI system, deriving from the product and processes of the system, from the trust given by a trustor, deriving from its "subjective estimation of the trustworthiness of an AI system" (Zhu et al. 2021, p. 3). Notably, the conceptualization of Zhu et al. (2021) also represents trust as a function of the trustworthiness of the trustee. Beyond conceptual rigorousness, this distinction is relevant for practice because of its tangible implications. It is to say that trustworthiness is a property of the system, and trust is an estimation of an agent; hence there can exist a mismatch, as pointed out by Zhu et al. (2021). The authors also identify some possible reasons for this, namely: "truster's numeracy (impacting the understanding of different types of trustworthiness evidence); a truster's prior beliefs and experiences; a truster's preferences and expectations on acceptable behaviors, types of evidence and explanation (Miller 2019); a system's observable behaviors to a truster." (Zhu et al. 2021, p. 4). Following the study of Miller (2019) on human-to-human explanation as a basis for explainable AI, a situation akin to a paradox can be presented here. Miller (2019), based on the idea that if the goal is to generate trust in a system, then a convincing explanation might be more effective than a precise explanation, there can be the case in which an explainer (developer or owner of an AI system) aims to generate trust and an explainee (human agent in this case) aims to just understand a decision. This is illustrative of a hypothetical, albeit realistic, situation when focusing on the explainability of an AI system to foster trust could lead to a non-accurate explanation. This has to be intended as an example of the potential -paradoxical- risks of a misunderstanding on the aspect of trust and trustworthiness.

The importance of trust in the post-adoption phases of technology is argued for in the literature. With a tailored approach, research also shows the relevance of trust in AI in its

use in the public sector. In particular, Kumar et al. (2021) show how trust is a fifth-order factor in the use of AI in post-distribution systems in the public sector in India. This means that trust underlies the use of the technology and also that other factors load on trust, making trust a fundamental element for using this specific AI in the case of post-distribution systems. In this context, the role of expectations in using AI by civil servants should also be recalled. This is crucial because, as discussed in the following subsection, literature on trust shows that trust is first and foremost a matter of expectation.

2.3.1 Trust in Artificial Intelligence

The concept of trust is relevant to a wide range of disciplines (Lee and See 2004; Robbins 2016; Thiebes et al. 2021). Due to its broad applicability, abundant literature focuses on it (Robbins 2016). Logically the interest of multiple disciplines in the concept of trust results also in multiple approaches to the concept as well as angles and levels of analysis. Whereas a synthesis of different conceptualizations of trust is out of the scope of this thesis, an overview of how trust is presented in the literature relevant to the topic of trust in AI is necessary. More so since some authors claim that trust pertains only to interpersonal relationships (Deley and Dubois 2020).

Generally speaking, scholars position specific trust in a nomological net where general dispositions to trust and institutional trust are the antecedents, trusting intentions are a consequence of the existence of certain conditions in the specific trust relationship, and trusting behavior is the result (Deley and Dubois 2020; Thiebes et al. 2021). Some scholars argue that the referents of trust are foremostly humans; hence trust is foremost a property of human relationships (Deley and Dubois 2020; Robbins 2016). For instance, building on Baier, Deley and Dubois (2020) claim that trust cannot exist in a human-machine relationship as the machine cannot betray the human but only fails to deliver the expected results, thereby becoming unreliable. Nonetheless, conceptualizations of trust in human-machine relationships are present in literature (Lee and See 2004; McKnight et al. 2011; Thiebes et al. 2021; Zhu et al. 2021). The McKnight-Chervany typology of trust categorizes concepts of trust into dispositional, institutional, and interpersonal (Deley and Dubois 2020). Disposition to trust relates to the reasons underlying the willingness to rely on others; institutional trust relates to the availability of structures that enable people that would not trust each other to do so; interpersonal trust relates to conditions existing between parties in a trust relationship, namely trusting intentions, beliefs, and behaviors (Deley and Dubois 2020). It should be noted that concepts of dispositional and institutional trust influence interpersonal trust in that they affect people's "intention to trust others, their beliefs of others' trustworthiness, and ultimately their trust-related behaviors." (Deley and Dubois 2020, p. 3). Additionally, it should be acknowledged that

this categorization of concepts of trust fits the nomological net of antecedents of specific trust, conditions of a trust relationship, and trust intentions referred to by Deley and Dubois (2020). Thiebes et al. (2021) draw a basic notion of trust from Mayer, which posits that trust is the willingness of a person to depend on another person because of the lack of total control that the former can exert on the latter. Robbins (2016) further defines the concept as a three-part relation where a trustor trusts a trustee for a specific matter at hand in a condition of unknown outcomes. In this thesis, grounded on these two latter conceptualizations of trust between people is the acceptance of trust in a human-machine relation. First, arguably humans do not hold complete control over technology and make themselves vulnerable to it (McKnight et al. 2011). Second, the definition of "matter at hand" encompasses "any resource, service or behavioral capability" of the trustee (Robbins 2016, p. 975). Therefore, it is argued here that such definition fits the use of technology made by humans, particularly so considering the matter at hand as a service (Robbins 2016). Third, whilst a situation of unknown outcomes (Robbins 2016) might not fit tools such as Microsoft Word, it may well fit more sophisticated technology that encompasses complex internal processes, such as AI (Lee and See 2004).

Notwithstanding this argumentation for the validity of a concept of trust in a human-machine relationship trust in people and technology differ in that the former reflects beliefs on the motives and attributes of other human beings (Thiebes et al. 2021). In contrast, the latter reflects beliefs regarding a technology's attributes as characteristics (Thiebes et al. 2021). However, scholars build on a similar argument to the one presented here and claim that the concept of trust is valid and relevant also concerning objects, specifically technology (Lee and See 2004; McKnight et al. 2011; Thatcher et al. 2011). On a foundational level, the conceptualization of trust in technology draws from interpersonal trust and is built on psychological and sociological studies (Lee and See 2004; McKnight et al. 2011). Additionally, trust in technology is embedded in the nomological net of trust recalled by Deley and Dubois (2020). Thus, it is influenced by those dispositions, structural elements, and conditions that are not solely related to the technology itself (Lee and See 2004; McKnight et al. 2011; Thatcher et al. 2011; Thiebes et al. 2021). Therefore, the trusting beliefs toward other humans and technology are related (McKnight et al. 2011; Thiebes et al. 2021).

Trust is essential in the relationship between humans and IT artifacts (McKnight et al. 2011; Thatcher et al. 2011). For instance, Thatcher et al. (2021) show how trust in technology influences the willingness of individuals to explore uses of technology after its adoption in the organization. Moreover, it is deemed particularly important to understand the relation between humans and a specific set of IT artifacts, namely autonomous systems (Lee and See 2004). Notwithstanding the formal distinction between

trust in IT artifacts and trust in autonomous systems, research shows how the two conceptualizations overlap (Thiebes et al. 2021). Based on their literature review, Lee and See (2004) claim that despite formal differences, the categories of the concept of trust in technology match theirs. It will become apparent through this section that this holds also in the case of categorizations following Lee and See's (2004).

Literature embeds the concept of trust in a nomological net (McKnight et al. 2011; Thiebes et al. 2021). On the first level are the dispositional concepts of trust identified in the McKnight-Chervany typology. In the case of trust in technology, general trusting beliefs are considered regarding technology (McKnight et al. 2011). Specifically, a favourable disposition towards technology consists of a general tendency to depend on a range of technologies across a spectrum of situations, the assumption that technologies are usually reliable, functional, and helpful for people, and the presumption that having a favourable disposition about technology one can achieve better outcomes (McKnight et al. 2011). The general disposition to trust technology is consistent with the findings of Lee and See (2004) regarding the influence of the individual and cultural context on trust. Institutional-based trust as a structural concept relates to the availability or absence of supportive structures and favourable situations (McKnight et al. 2011). The institutional-based trust consists of the belief that the use of a specific technology is likely successful because the user already feels comfortable in using the generic type of technology of which the specific one is an instance and because the given technology is embedded in a system of support, safety and guarantees mechanisms (McKnight et al. 2011). Institutional-based trust also has people as referents. Zhu et al. (2021) argue that it is important to distinguish product and process mechanisms of assurance and evidence of trustworthiness in operationalizing responsible AI. Regarding trust, this means that whether people perceive processes and structures as trustworthy influences their trust in the AI system (Zhu et al. 2021). It is to say that the institutional-based trust consists of a double referent. The human referent of institutional-based trust is consistent with Lee and See's (2004) categorization of concepts of trust in that they find that organizational context influences trust in autonomous systems. McKnight et al. (2011) present trust in a specific technology as trust in IT artifacts. As anticipated, Lee and See (2004) present it as trust in autonomous systems. Thiebes et al. (2021) show the overlap between the categorizations proposed by these authors. The categories of performance, purpose, and process introduced by Lee and See (2004) somewhat match the categories of functionality, helpfulness, and reliability introduced by McKnight (2011). Nonetheless, it should be highlighted here the differences. The category of purpose shows an important difference from the category of helpfulness as the former refers to the extent of match between the actual use of technology by the operator and the use intended by the designer, whereas the latter relates to accurate and responsive help to the user (Lee and See 2004;

McKnight et al. 2011). However, the help that the technology offers to the user is based on the designer's intended use in that the help function is part of the development. Additionally, process refers to the characteristics of the autonomous system and whether these are understood and deemed appropriate to the user's goals, whereas reliability relates precisely to the perceived consistency of the operations carried out by the technology (Lee and See 2004; McKnight et al. 2011). Nonetheless, there is a connection between the understanding of the autonomous system by the user, a positive evaluation of its appropriateness, and the belief that the technology consistently operates properly. These two categorizations of trust in a specific technology are consistent with the trust in the product highlighted by Zhu et al. (2021). Importantly in this regard is that the category of the process of the autonomous system introduced by Lee and See (2004) refers to the internal process, thereby being different from the ensemble of processes surrounding the AI system as categorized by Zhu et al. (2021), mentioned before as a referent of institutional-based trust. As a final remark, it should be noted that, for this thesis, the focus is on the civil servants' perceptions, beliefs, and expectations concerning the concepts mentioned so far. It is to say that, as argued above, despite the development of a trustworthy system, the decision on whether to trust it or not is ultimately dependent on the belief of the trustor about the trustworthiness. It follows that, for instance, following Zhu et al.'s (2021) distinction between product and process and people, the development of explainable AI increases trust only if the truster considers the explanation of the AI system as understandable and aligned with its expectations or, alternatively, if a process of stakeholders representing the truster and assuring the explainability of the AI system is in place and deemed trustworthy by the truster. Additionally, albeit important, a holistic approach to trustworthiness moves to the background since Robbins' (2016) concept of trust as a tripartite relation is of utter importance, and its validity in the realm of trust in technology is shown by literature (Aoki 2020).

Given the account of the extant literature on trust and trust in technology made so far, conclusions can be drawn regarding the validity of the concept of trust in technology. This thesis considers a combination of conceptualizations: trust in a specific technology (McKnight et al. 2011) and trust in automation (Lee and See 2004). In addition, the focus on people in the concept of trust in Artificial Intelligence brought forward by Zhu et al. (2021) is included. Therefore, the conceptualization of trust used in this thesis is framed as trust in Artificial Intelligence, aside from when a specific element of one of the abovementioned conceptualizations is mentioned. The concept of trust in Artificial Intelligence, therefore, consists of the following dimensions. First, is the dimension of a general disposition to trust technology (McKnight et al. 2011). This consists of a belief that using technology leads to better outcomes than refraining from its use (McKnight et al. 2011). In other words, it consists of a tendency to trust technology (Lee and See 2004).

Second, is the dimension of institutional-based trust (McKnight et al. 2011). This dimension consists of a belief that the use of a specific technology is likely successful because the user already feels comfortable in using the generic type of technology of which the specific one is an instance and because the given technology is embedded in a system of support, safety and guarantees mechanisms (McKnight et al. 2011). This system includes organizational structures and culture (Lee and See 2004). In this sense, interpersonal trust is an element of this dimension (Zhu et al. 2021). Third, is trust in the specific technology (McKnight et al. 2011) -i.e., Artificial Intelligence systems as defined for this thesis in subsection 2.1. This dimension consists of the elements of functionality, reliability, and internal processes (Lee and See 2004; McKnight et al. 2011). In other words, whether an AI system is deemed functional for the tasks of the human agent or the public organization at large, reliable, and its internal processes are considered understandable determines trust in the AI system (Lee and See 2004; McKnight et al. 2011).

To summarize, extant research on the impact of the use of AI in the public sector builds on previous efforts made by scholars to assess the impact of technology on governance (Bullock et al. 2020; Lorenz et al. 2021; Young et al. 2019). Overall, the literature review highlights that the impact of the use of AI technologies by public sector organizations on the government is ubiquitous and difficult to determine (see subsection 2.1). This conclusion underlies the efforts of scholars studying the factors influencing the adoption and post-adoption phases of AI in the public sector (Pencheva et al. 2020; Van Noordt and Misuraca 2020; Veale and Brass 2019). Consistently with the socio-technical approach to the study of technology, the results of the use of AI by public organizations is influenced by multiple factors (Bailey and Barley 2020; Guenduez et al. 2020; Meijer et al. 2021). Research on the junction between AI adoption and post-adoption phases in the public sector increasingly focuses on individual-level factors (Ahn and Chen 2021; Alshahrani et al. 2021; Kumar et al. 2021). Notably, the relevance of factors influencing the adoption and use of AI in the public sector on an individual level includes public sector employees at all levels of government (Ahn and Chen 2021; Alshahrani et al. 2021; Sun and Medaglia 2019; Veale and Brass 2019). This focus is due to the relevance of subjective perceptions from, for instance, the decision to explore potential uses of AI to its introduction in public organizations and evaluation of its impact on public value creation (Ahn and Chen 2021; Alshahrani et al. 2021; Van Noordt and Misuraca 2020; Veale and Brass 2019). In this regard, trust in AI has captured the attention of both scholars and policymakers (Thiebes et al. 2021). Significantly, subjective perceptions also influence the trust that human agents have in a technology -hence, AI- (Lee and See 2004; McKnight et al. 2011; Zhu et al. 2021). It is argued here that, whilst stressing the importance of subjective perceptions and trust, severe shortcomings on this matter

influence both research and practice. On the one hand, extant research on the subjective perceptions of civil servants on AI primarily focuses on challenges specifically related to the adoption and diffusion of AI (Alshahrani et al. 2021; Sun and Medaglia 2019). Scholars who seek to fill this gap do so by focusing on general perceptions about AI technologies (Ahn and Chen 2021). On the other hand, research on trust in AI suffers from conceptual imprecision (Zhu et al. 2021). Indeed, trustworthiness is placed at the front, although this is a property that AI systems may or may not have (Zhu et al. 2021). A human agent may or may not trust AI based on his or her estimate of the system's trustworthiness (Zhu et al. 2021). As a result, also policymakers focus on the trustworthiness of AI (Thiebes et al. 2021), thereby overlooking the trustor in the trust relationship and focusing solely on the trustee (Zhu et al. 2021). Therefore, this thesis seeks to contribute to the literature by examining subjective perceptions of civil servants on the use of AI in the public sector with a focus on trust. It does so by building on the concepts of trust in a specific technology (McKnight et al. 2011) and trust in automation (Lee and See 2004). Such concepts are used to frame the subjective perceptions of civil servants on AI. In so doing, the subjectivity of civil servants on the use of AI in the public sector is explored by putting the subjective perceptions of AI concerning the trust in AI. This constitutes the conceptual framing of the viewpoints of Estonian civil servants on the use of AI in the Estonian public sector.

3 Research Approach and Methodology

3.1 Q-Methodology

E-government is a relatively new field of research that would benefit from a compelling accumulation of knowledge (Heeks and Bailur 2007). Research should be positioned against its philosophical underpinnings to contribute to a sound accumulation of knowledge (Heeks and Bailur 2007). A clear philosophical stance benefits the accumulation of knowledge in that it clarifies the researcher's assumptions on the very origin of knowledge and the practices needed to create it (Gregor 2006; Heeks and Bailur 2007). In doing so, each contribution can be positioned in the debate in the field and provide the tools to evaluate it in contrast and comparison to other contributions that move from different philosophical stances (Heeks and Bailur 2007).

Building on this, the philosophical stance adopted in this thesis is briefly explained here, whereas the following sections discuss the specificities of the methodology used for this research. The present thesis moves from a critical-realistic stance. Reducing critical realism to its core, this philosophy combines a realist ontological perspective with a certain degree of epistemological relativism (Zachariadis et al. 2013). This twofold perspective on ontology and epistemology finds its underpinnings in the twofold view of reality and knowledge (Zachariadis et al. 2013). Zachariadis et al. (2013) draw on the founding father of critical realism -i.e., Bhaskar- and illustrate how such philosophy conceives reality as existent independently from human knowledge or perception and knowledge as a human activity is socially produced. Seemingly convoluted, the dual view of ontology and epistemology proper of critical realism is perhaps best explained by the layered view of reality put forward by this philosophy. In this sense, the synthetic account proposed by Zachariadis et al. (2013) suffices for this section. Zachariadis et al. (2013) retrace the critical-realistic ontology posited by Bhaskar in the three dimensions of real, actual, and empirical. “The domain of the real includes objects and structures with inherent causal powers and liabilities which result in mechanisms that may not be visible.” (Zachariadis et al. 2013, p. 857). The actual consists of a subset of the real, which includes events generated by said mechanisms (Zachariadis et al. 2013). Lastly, the empirical consists solely of the phenomena that can be observed (Zachariadis et al. 2013). The ontological view posited by critical realism appears consistent with the vital relationship at the center of this thesis. Specifically, this is the distinction Zhu et al. (2021) highlight between trustworthiness as the property of a system and trust as the estimation of a human agent of the trustworthiness of such system. Building on this distinction, it can be argued that the existence of certain elements that qualify a system as trustworthy is independent from their perception by a human agent. However, the trust given by a

human agent to a system is ultimately dependent on the perception of those elements by said human agent. Accordingly, Zhu et al. (2021) highlight this distinction shedding light on the potential mismatch between trustworthiness and trust. Drawing on the ontological view, critical realism posits that different research methods have different roles in the epistemological endeavor of researchers (Zachariadis et al. 2013). Namely, quantitative methods have a descriptive role, whereas qualitative methods are more suited to identify underlying relationships and account for them (Zachariadis et al. 2013). In this sense, mixed-method approaches are the solution to epistemological fallacies arising from using a single-method approach (Zachariadis et al. 2013). Therefore this research adopts a mixed-method approach.

The mixed-method chosen for this thesis is Q-methodology. Q-methodology was first introduced by William Stephenson (Brown 1993). Confined in the realm of psychology, Q-methodology began to find consensus outside of such field only a few decades ago (Brown 1993). The reason appears to be found in the realization that Q-methodology "provides a foundation for the systematic study of subjectivity" (Brown 1993, p.93). In other words, Q-methodology consists of a systematic study of subjective perspectives of individuals and results in the elicitation of ideal-types of thought as they appear in the population under study (Brown 1993). Recently, Q-methodology was used in research particularly close to this thesis. Guenduez et al. (2020) apply Q-methodology to the study of technological frames of public managers on big data. Significantly the authors refer specifically to the fitness of Q-methodology with the study of subjectivity to justify its use in the study of technological frames (Guenduez et al. 2020). In this thesis context, Q-methodology appears to be the best fit for a similar reason. This research draws on Zhu et al. (2021) conceptual distinction between trustworthiness and trust: the former being a property of the system, the latter being an estimate of the trustworthiness by a human agent. In this context, the beliefs and expectations of human agents become prominent factors in shaping their estimate of the trustworthiness of AI (Zhu et al. 2021). Therefore, the subjectivity of civil servants constitutes a prominent factor regarding their estimate of the trustworthiness of AI systems. As a result, studying the subjectivity of civil servants regarding Artificial Intelligence is the objective of this research. Therefore, Q-methodology constitutes the most suitable option because it provides a systematic approach to the study of said subjectivity and renders ideal-types of thought as they appear among civil servants (Brown 1993). Additionally, Q-methodology applies quantitative methods in that it relies on factor analysis but assigns the qualitative interpretation of factors the primacy in interpreting results (Brown 1993; Dziopa and Ahern 2011). This use of both quantitative and qualitative methods is consistent with the prerogatives of critical-realism. Additionally, as will be exposed in this section, the function of the two methods in Q-methodology is consistent with the role that critical

realism posits for them. Q-methodology's ultimate value is eliciting different configurations of the elements proper of a given subject as they appear among the study participants (Watts and Stenner 2005). Q-methodology relies on a fundamental inversion of traditional factor analysis (Watts and Stenner 2005). The study sample consists of the measurable items -i.e., statements on a given subject- whereas the variables consist of the individuals (Watts and Stenner 2005). Factor analysis in Q-methodology puts the statements on a given subject -i.e., the measurable items- in relation to each other, and the individuals -i.e., the variables- are loaded onto the factors (Watts and Stenner 2005).

3.1.1 Concourse and Q-Sample

Q-methodology starts with the development of the concourse (Brown 1993; Watts and Stenner 2005). The concourse is the "flow of communicability surrounding any topic" (Brown 1993; p. 95). In other words, the concourse represents the broad spectrum of opinions surrounding any topic (Watts and Stenner 2005). In this regard, the preferred development of the concourse is based on interviews conducted with a sample of the population under study (Brown 1993). This approach is preferred because the right level of sophistication is sought in developing the concourse and the Q-sample afterward (Brown 1993). The concourse for this thesis aims to represent the broad debate surrounding the use of Artificial Intelligence in the public sector, focusing on trust in Artificial Intelligence systems. Therefore, resources to develop the concourse are gathered with this objective. Specifically, considering the time constraints of this research, resources are gathered that include the perspective of public sector employees and key stakeholders on the use of Artificial Intelligence in the public sector. These may be, among others, research including interviews with public sector employees at all levels, keynotes and panels of sector-specific conferences, and publications of international organizations or private organizations specifically targeted at the public sector. Given the resources available for this thesis, developing the concourse through interviews is deemed unfeasible for time constraints. Instead, scientific papers, sector-specific conferences, and documents gathered through the researcher's network are used to develop the concourse. A list of such material is available in Appendix A. It appears necessary to clarify that the sources used are encompassed by Q-methodology as using interviews is not a strict requirement but rather a preferred approach (Brown 1993; Watts and Stenner 2005).

The next step after the concourse development is extracting a representative sample of statements, the so-called Q-sample (Brown 1993). Importantly, the construction of the Q-sample is a critical passage in Q-methodology because it consists of the moment when the broad spectrum of opinions is reduced to a set of statements that are meant to represent it and is small enough for practical reasons (Brown et al. 2018; Paige and Morin 2014).

The Q-sample can be constructed using inductive and deductive approaches (Paige and Morin 2014). The construction of a Q-sample based on a deductive approach is commonly referred to as structured (Brown et al. 2018; Paige and Morin 2014). The structured construction of a Q-sample consists of reducing the statements of the concourse to a sample that represents its diversity based on theoretical constructs or concepts (Brown et al. 2018). With regards to this, albeit valid regardless of the approach to constructing the Q-sample, it becomes clear that the research goal or question is pivotal (Watts and Stenner 2005). Therefore, the concept of trust in technology becomes central in the construction of the Q-sample for this thesis. The works of Paige and Morin (2014) and Brown et al. (2018) offer practical guidelines for the structured construction of a Q-sample. These were thoroughly followed for the construction of the Q-sample for this thesis. Accordingly, the concourse statements were categorized based on the concept of trust in Artificial Intelligence. This led the operationalization of the concept of trust, which in turn determined the selection of statements. The conceptual map of trust in technology is available in Appendix B. It is important to note that trust in technology is embedded in the nomological net of trust (Thiebes et al. 2021). In the context of the development of the Q-Sample, this implies the inclusion of statements that relate to general dispositions to trust technology; to institutions and structures that enhance trust when this may not exist otherwise; to specific characteristics of technology (Lee and See 2004; McKnight et al. 2011; Zhu et al. 2021). In addition to the object of the statements referring to said dimensions, statements were considered to refer to the concept of trust if they expressed opinions consistent with the conceptualization of such dimensions found in the literature. It is to say that statements on the dimension of general disposition to trust technology were included if they expressed an opinion on whether using technology leads to better or worse outcomes than not (McKnight et al. 2011). Statements on the dimension of institutional-based trust were selected if they expressed an opinion about a system of supporting and safeguard mechanisms (McKnight et al. 2011), including people other than structures (Zhu et al. 2021) and organizational context and culture (Lee and See 2004). Statements on the dimension of specific trust in Artificial Intelligence were included if they expressed an opinion about the functionality, reliability, or internal processes (Lee and See 2004; McKnight et al. 2011). Afterward, statements were selected based on the two criteria of homogeneity and heterogeneity (Brown et al. 2018). It is to say that statements were first selected homogeneously in each category, thereby including statements in each category based on their similarity. For instance, statements addressing the usefulness of technology for human wellbeing and statements addressing the environmental impact of computation were categorized as referring to general dispositions towards technology (Lee and See 2004; McKnight et al. 2011). Similarly, statements addressing factors identified by the literature as influencing the use of

Artificial Intelligence in the public sector on the institutional and organizational levels were categorized as referring to the institutional-based dimension of trust in technology (Lee and See 2004; McKnight et al. 2011). Finally, statements addressing specific aspects of Artificial Intelligence technologies were categorized as referring to the dimension of trust in a specific technology (McKnight et al. 2011) -trust in automation (Lee and See 2004). Second, statements within categories were selected heterogeneously, thereby including statements on each category that express different opinions on that aspect. With this regards it shall be re-stated here the purpose of Q-sampling, hence representing the diversity of the opinions expressed in the concourse (Brown et al. 2018). Additionally, each statement must contain only one proposition (Watts and Stenner 2005) and express an opinion and not a fact (Brown et al. 2018). For these reasons statements should be-reworked in the process of construction of the Q-sample (Brown et al. 2018). Statements were reworked so as to make them less sophisticated. However, this process is admittedly open to fallacies, thereby not granting that the level of sophistication of the statements is fully adequate to the population under study. Following this procedure, a Q-sample consisting of thirty-nine statements is constructed. The Q-Sample is available in Appendix C. Open questions interviews have therefore been used to contextualize respondents' opinions, thereby enriching the findings of this thesis. Q-Methodology includes a specific research strategy, namely Q-Sorting (Brown 1993; Watts and Stenner 2005). This is addressed in the next sub-section.

3.1.2 Q-Sort

The research strategy for research methodology is defined explicitly as Q-sorting (Brown 1993). Q-sorting consists of the following steps. Participants are given a set of statements -i.e., the Q-sample- in random order and asked to rank the statements on a scale (Watts and Stenner 2005). Afterward, participants are given the option to re-arrange the statements if they wish to do so (Watts and Stenner 2005). The scale used in this thesis is a seven-point scale -i.e., -3/+3. A particular, often criticized, feature of Q-methodology is that participants are asked to assign a rank in a fixed quasi-normal distribution (Watts and Stenner 2005). Whereas the use of a fixed distribution -also called forced Q-sort- has been often criticized, scholars prove that "the chosen distribution actually makes no noticeable contribution to the factors which emerge from a particular study" (Watts and Stenner 2005, p. 77). The fixed quasi-normal distribution also determines the number of slots available in each ranking position, hence the number of statements that can be assigned that rank (Watts and Stenner 2005). A representation of the fixed quasi-normal distribution is included in Appendix D. Q-sorting is completed by gathering supporting information, so-called post-sorting questions (Watts and Stenner 2005). Such information is gathered through open-ended questions and is most valuable when they revolve around

the statements ranked the highest, the lowest, in the middle of the scale, and any other statement that the participant would like to comment on (Watts and Stenner 2005). Additionally, questions regarding the AI project the participants are involved in are added to this thesis to gather further helpful information to contextualize Q-sorts. Contextualized interpretation, indeed, is the ultimate goal of post-sorting questions (Watts and Stenner 2005). The selection of respondents -i.e., the definition of the P-sample- for Q-methodology should not aim to be representative of the population under study but rather aim for heterogeneity (Watts and Stenner 2005). Doing so increases the possibility of eliciting unexpected configurations (Brown 1993). Eliciting otherwise overlooked configurations of ideas is, ultimately, the main goal of Q-methodology (Brown 1993). For this reason, purposeful sampling is preferred to representative sampling for this thesis.

Q-sorting is conducted digitally. Considerations on the benefits of using a digital tool to help participants visualize their ranking -i.e., the fixed quasi-normal distribution- led to choosing Miro. Miro is an online collaborative whiteboard that can be used according to the user's needs. Using Miro also allows the researcher to see the sorting in real-time. For this thesis, statements are written on digital post-its, and the participants are asked to position them on a frame representing the fixed quasi-normal distribution. First, participants are asked to order statements one by one, with the possibility of deviating from the fixed distribution. Subsequently, they are given additional time to re-order them respecting the fixed distribution. Participants are encouraged to think out loud while sorting the statements to help the researcher follow their reasoning and add meaningful context to the final Q-sort. Q-sorting is conducted over Microsoft Teams meetings. The use of Microsoft Teams allows for automatic transcription and recording of the data collection. However, one participant bound participation to the possibility of working autonomously and answering the post-sorting questions in a written format. The participant was allowed to do so in order to include as many participants as possible. Each meeting lasted one hour, of which roughly fifteen to twenty minutes were devoted to post-sorting questions, depending on the time spent by participants sorting the statements.

The development of the P-Sample and the resulting selected cases are the objects of the following sub-section.

3.2 P-Sample

The definition of the P-sample for this thesis is conducted as follows. The Estonian Artificial Intelligence Strategy for 2019-2021 is consulted. Information on the use-cases of Artificial Intelligence in Estonia is gathered from said strategy. Information on said use-cases is openly accessible as they are available online in English (see: en.kratid.ee/kasutuslood). Based on the information available, cases are selected that are

related to Artificial Intelligence systems that fit under the definition of Artificial Intelligence systems in the scope of this thesis presented in the theoretical background section. Use-cases in different public services fields are sought to account for different public services domains, recalling Vydra and Klievink (2019) findings, among others. After preliminary research, four use-cases are identified that match the requirements. The cases selected are use-cases of Artificial Intelligence in the following fields: emergency services, education, social affairs, and healthcare. After identifying the first point of contact for each use-case, the snowball sampling technique is used (Naderifar et al. 2017). Snowball sampling is chosen to allow some leeway in constructing the P-sample. Therefore, the first contact points are asked to provide information regarding other civil servants involved in the AI project with specific characteristics. Such characteristics are limited to the role of said civil servants in the AI project at hand or in the public organization where the AI project is conducted, recalling among others Veale and Brass (2019). These roles are the senior manager responsible for the information systems in the organization -this can be the Chief Data Officer, the Chief Information Officer, the Chief Technology Officer, or else-; the project manager; and at least one and up to three users, potential users, or project team members -these can be first-level bureaucrats, data analysts, or else. Following this definition of the P-sample, this shall consist of twelve to twenty participants. The P-sample should be composed of four senior-level managers, four project managers, and four to twelve participants among users, potential users, or project team members. The P-sample should consist of three to five civil servants for each use-case identified. Therefore, it should be composed of three to five civil servants for each public service domain where said use-cases are identified -emergency services, social affairs, education, and healthcare.

3.2.1 Selected Cases

Enhanced Care Management – Healthcare

The Enhanced Care Management is an AI project developed through the collaboration between the Estonian Ministry of Health, the World Bank Organization, and a private organization. This AI project aims to improve primary care in Estonia by identifying individuals at risk and supporting general practitioners with diagnosis and therapy. This AI project developed an AI system that learns from health data on the Estonian population through Machine Learning. The AI system aims to predict at-risk individuals who could benefit from primary healthcare. Subsequently, the AI system supports general practitioners in the diagnosis. This AI system for primary healthcare is already implemented in Estonian healthcare.

Personalized Learning and Teaching – Education

The personalized learning and teaching project is an in-house AI project developed by the Estonian government. The peculiarity of this AI project is that it is being developed entirely in-house and is an open-source project. This AI project aims to develop a suite of tools that supports students and teachers. The goal of the AI system powering the tools is to advance learning analytics by analyzing learning data and allowing the personalization of learning and teaching. The AI project is ready to be piloted. However, learning data is sensitive, and authorities closely supervise its use due to the GDPR. Hence, the project is not piloted yet.

The Threat Assessment Helper of the Emergency Response Centre – Emergency Services

The Threat Assessment Helper of the Emergency Response Center is an AI project in the field of emergency services developed by the Estonian government and a private organization. The Estonian government was involved through both the Emergency Response Center and the IT and Development Center of the Ministry of Interior. This AI project developed an AI system that learns from emergency calls through Machine Learning. The AI system listens to and transcribes calls from car accidents while analyzing various data, such as environmental noise and voice tone. The goal of the AI system is to recommend the dispatcher through a risk assessment of the emergency. The AI system was developed and successfully piloted in 2020. However, the AI system is not implemented.

OTT – Social Affairs

OTT is an AI project developed in the field of social affairs by the collaboration between the Estonian government, academia, and a private sector organization. This AI project aims to improve the services to unemployed individuals. This AI project developed a system that learns from population and labor market data through Machine Learning. The goal of the system is to forecast the risk of an individual being long-term unemployed and support consultants. The AI system has been used since 2021 at the Estonian Unemployment Insurance Fund.

Some changes from the original P-sample became necessary. These changes regarded the number and role of participants in each selected case. Nonetheless, this was somewhat foreseen, and this consideration led to the choice of snowball sampling techniques in the first place. The P-sample consisted of 15 participants. In contrast to representative sampling techniques, purposeful sampling techniques do not require the sample to be representative of the population under study. In fact, within the context of Q-methodology, purposeful sampling aims to include participants with different backgrounds to consider different perspectives on the subject under study. The rationale

being eliciting insightful and overlooked configurations of ideas on the matter at hand. Participants are civil servants in several Estonian public organizations. The policy areas included in this research are the following: healthcare, education, emergency services, and social affairs. The role of participants within their organization and projects involving AI range from senior-level managers to street-level bureaucrats. In light of the purposeful sampling technique, the number of participants per policy area and per role included in the P-sample shall not be considered representative of the population. Following the data collection phase, one participant was excluded as its Q-sort was unsuitable for data analysis, reducing the P-sample to 14 participants.

Table 1 provides an overview of the P-Sample. Each participant was assigned a unique code -i.e., ID- consisting of the country code of Estonia -i.e., EE- and a number corresponding to the order in which the participant was interviewed. Therefore, participants will be referred to in the rest of this dissertation as EE1, EE2, EE3, until EE14. The code of the participant excluded from the data analysis is EE0. Additionally, Table 1 reports the characteristics of participants selected for the purposeful sampling -i.e., the public organization where participants work; the related public service domain; and their role, divided into primary and secondary depending on whether participants mentioned a secondary role in addition to their job title.

PUBLIC ORGANIZATION	**PUBLIC SERVICE DOMAIN**	**PRIMARY ROLE**	**SECONDARY ROLE**	**ID**
Haigekassa	Health		N/A	EE0
Hairekassa	Emergency	Information Manager	N/A	EE1
Haigekassa	Health	IT Manager	Integration manager	EE2
Haigekassa	Health	Subject Expert	N/A	EE3
Haigekassa	Health	Project Manager	N/A	EE4
Ministry of Education	Education	Project Manager	N/A	EE5
Ministry of Education	Education	Head of Workshop	Product Owner	EE6
Clinic	Health	GP	Expert	EE7
Harno	Education	Head of Innovation	N/A	EE8
Clinic	Health	GP	Expert	EE9
High School	Education	Teacher	N/A	EE10
High School	Education	Teacher	N/A	EE11
Tootukassa	Social	User Trainer	Representative	EE12
Tootukassa	Social	Analyst	Integration officer/Change Manager	EE13
High School	Education	Teacher	N/A	EE14

Autor's own elaborate

Tab. 1 P-Sample Overview

Data collected through the Q-Sorting and post-sorting questions are subsequently analyzed according to the Q-Methodology. This consists of factor analysis and interpretation (Watts and Stenner 2005). This phase is the object of the following sub-section.

3.3 Data Analysis

Data analysis for Q-methodology consists primarily of factor analysis (Brown 1993; Watts and Stenner 2005). Secondarily, factors are interpreted through the qualitative analysis of data (Brown 1993; Watts and Stenner 2005). Factor extraction is performed using the centroid method. The use of this method for factor extraction in Q-methodology is suggested because of its indeterminacy in the extraction of factors (Brown 1993). Factor extraction with the centroid method is performed using Brown's centroid method for factor extraction. Subsequently, factor rotation is performed using the theoretical (or judgmental) rotation. The founding fathers of Q-methodology argue for the use of theoretical (or judgmental) rotation (Brown 1993; Watts and Stenner 2005). The main argument is that theoretical rotation relies on the researcher's knowledge about the research subject and allows insights derived from the post-sort questions to be explored (Brown 1993). The rationale of this argument is that Varimax maximizes the variance explained by the factors from a purely statistical (mathematical) perspective (Watts and Stenner 2005). However, meaningful explanations with regards to the matter at hand are not necessarily statistically relevant (Brown 1993; Watts and Stenner 2005). For example, the viewpoint of one single participant could be particularly relevant for the research and worth be matched with a factor to explore its position in comparison to other participants' positions (Watts and Stenner 2005). This could be the case if the researcher is interested in, say, a CIO's viewpoint on AI in comparison to developers, program, and project managers with regards to the adoption of AI in an organization in the context of agenda setting and an AI strategy. Even though this is not the case with this thesis. It is worth noting that theoretical rotation does not equal theoretical certainty or validation of theory (Brown 1993; Watts and Stenner 2005). In other words, the purpose of the theoretical rotation is to explore theoretically relevant insights as they emerge from the researcher's view (Brown 1993; Watts and Stenner 2005). Ultimately, eliciting and highlighting such insights is somewhat of a prerogative of the Q-methodology (Brown 1993; Watts and Stenner 2005). However, this first step shall be followed by further analysis and explanatory research (Brown 1993; Watts and Stenner 2005). This, one could argue, is why an alternative naming of theoretical rotation is judgmental rotation. The ultimate output of the quantitative component of the Q-methodology is the discovery of factors across the Q-sorts (Brown 1993; Watts and Stenner 2005). The composite Q-sort of a factor is its representative Q-sort (Brown 1993). The composite Q-sort derives from the

Q-sorts that load onto a given factor, thereby constituting the representative score that a factor gives to each statement, hence the factor scores (Brown 1993). Such factor scores ultimately determine the similarity of the viewpoints expressed by the factors (Brown 1993). In different words, factor scores indicate to what extent the viewpoints of the factors agree or disagree with the statements of the Q-sample, thereby determining the relative position of their views on the matter under study (Brown 1993; Watts and Stenner 2005). Factor analysis in Q-methodology relies on the scores given by participants in the Q-sort to extract the factors (Brown 1993; Watts and Stenner 2005). Importantly, the interpretation of the results of the quantitative analysis follows primarily the factors scores rather than the factor loadings (Brown 1993). It is to say that the focus of the interpretation of the composite Q-sorts is the factor scores, hence the representative score given by a factor to each statement. Arguably, a peculiarity of the Q-methodology is that it aims for a deeper understanding of the configurations of ideas than what emerges from the quantitative analysis of the Q-sorts, and is validated by statistical significance (Brown 1993; Watts and Stenner 2005). In fact, it enables the researcher to draw from quantitative analysis and go beyond numbers, analyzing the factors scores in combination with the post-sort questions (Brown 1993; Watts and Stenner 2005). In the words of Watts and Stenner (2005, p. 82): "The interpretative task in Q methodology involves the production of a series of summarizing accounts, each of which explicates the viewpoint being expressed by a particular factor. These accounts are constructed by careful reference to the positioning and overall configuration of the items in the relevant 'best-estimate' factor arrays. The obvious place to begin such deliberations is at the two 'poles' of the Factor configuration". It is to say that the consideration of the composite Q-sorts is pivotal to the interpretation of results in Q-methodology. Factor interpretation shall consider the two poles of the composite Q-sorts -i.e., +3 and -3 in this research- and the neutral area -i.e., 0 (Brown 1993; Watts and Stenner 2005). For a comprehensive account of the holistic nature of the configurations, their interpretation should not be limited to the poles and the center of the composite Q-sorts and should include the answers to the post-sort questions as well (Watts and Stenner 2008). The goal of the interpretative task of Q-methodology is precisely rendering such a holistic interpretation of the factors' configurations (Brown 1993; Watts and Stenner 2005). The quantitative analysis can only highlight distinguishing statements from a statistical point of view (Brown 1993; Watts and Stenner 2005). Therefore, solely relying on the quantitative analysis exposes the researcher to the risk of focusing on too few of the items of the Q-sort (Brown 1993; Watts and Stenner 2005). A narrow focus would result in a severe interpretative failure given the purpose of Q-methodology (Brown 1993; Watts and Stenner 2005). As a result, the qualitative analysis of post-sort answers gathered from participants who load significantly on each factor becomes crucial for correctly interpreting results on the configurations (Brown

1993; Watts and Stenner 2005). Putting this in the perspective of the ontological stance of critical-realism adds to the contextualization of the research methodology from a research philosophy's point of view. Indeed, the quantitative analysis of data fulfills descriptive purposes, whereas the qualitative analysis of data constitutes the core of the interpretation of data conducted by the researcher (Brown 1993; Watts and Stenner 2005; Zachariadis et al. 2013).

Data analysis is conducted using KADE, a desktop application for Q-Methodology. KADE is chosen because it is open-source software and because of its features. Specifically, KADE includes interactive visualizations and a graphic user interface that facilitates the user's understanding and manipulation of data (Banasick 2019). Data collected through Miro was transcribed on a .xlxs file consistently with the requirements of KADE's data entry specifics. This consisted of listing all the statements and their numbers and reporting the participants' sortings listing the number of the statement and its rank. Afterward, the .xlxs file was uploaded onto KADE for the data analysis. A .xlxs file report of the data analysis is available in the supporting materials.

The detailed description of the data analysis and its results is the object of the next section.

4 Results

4.1 Factor Analysis

In Q-methodology, the correlation matrix is deemed a necessary passage for the data analysis and the extraction of the factors' structure in that the correlation matrix reports the (dis)similarity of the participants' perspectives (Brown 1993). However, the attention is focused on the factors on which the correlation leads. Nonetheless, the correlation matrix provides insights into the homogeneity or heterogeneity of perspectives captured in the P-sample and highlights significantly similar perspective pairs (Brown 1993). According to Brown (1993), a rule of thumb to determine the threshold for statistically significant correlations is: a correlation is considered statistically significant if it is approximately 2 to 2.5 times the standard error, irrespective of the sign. Hence, correlations are considered statistically significant in this case if they are between 0.32 and 0.4. As reported in Table 2, the highest correlation is observed between the perspectives expressed by participants EE5 and EE10. This is consistent with some similar characteristics of the two participants, namely the policy area and the role in their organizations. They both work in the field of education. Additionally, EE5 is a project manager in a specific project for developing AI systems in the field of education that admittedly works in close contact with potential users, hence teachers. EE10 is a teacher responsible for using technology in education in its school. Perhaps more interestingly, the second highest correlation is between EE6 and EE12. In this case, the two participants differ in those same characteristics. They work in different policy areas, namely education and social affairs, and their roles are almost at the two ends of a hypothetical continuum between strategy-making and the use of technology. Indeed, EE6 is a program manager, whereas EE12 is a user trainer and representative.

	EE1	EE2	EE3	EE4	EE5	EE6	EE7	EE8	EE9	EE10	EE11	EE12	EE13	EE14
EE1	100	28	-8	10	3	33	11	23	30	-6	-8	32	-15	8
EE2	28	100	29	26	6	27	3	16	33	35	11	41	9	35
EE3	-8	29	100	5	16	16	-13	19	10	10	27	13	8	48
EE4	10	26	5	100	25	21	43	24	25	33	-3	38	18	12
EE5	3	6	16	25	100	32	-8	7	26	59	-14	39	27	31
EE6	33	27	16	21	32	100	-9	21	25	26	-6	52	19	34
EE7	11	3	-13	43	-8	-9	100	14	12	-1	21	8	-2	-24
EE8	23	16	19	24	7	21	14	100	3	-3	-7	8	-15	16
EE9	30	33	10	25	26	25	12	3	100	47	30	41	48	48
EE10	-6	35	10	33	59	26	-1	-3	47	100	17	47	43	36
EE11	-8	11	27	-3	-14	-6	21	-7	30	17	100	-7	17	22
EE12	32	41	13	38	39	52	8	8	41	47	-7	100	44	23
EE13	-15	9	8	18	27	19	-2	-15	48	43	17	44	100	12
EE14	8	35	48	12	31	34	-24	16	48	36	22	23	12	100

Author's own elaborate

Tab. 2 Correlation Matrix

Besides highlighting particularly significant correlations between the perspectives expressed by pairs of participants from a statistical standpoint, there is little value in commenting on the correlation matrix in Q-methodology. Of higher value for the objective of the research and the data analysis is the factor extraction from the correlation matrix. The software used for the data analysis automatically performs a 7-factors factor extraction; however, a 3-factors factor extraction was chosen for this dataset.

Factor extraction resulted in the extraction of three factors. All three factors were kept for rotation. Factor analysis relies on eigenvalues to determine the factors for factor rotation. Namely, factors with eigenvalue > 1 explain the variance in the dataset higher than the average, thereby being worth rotation. Nonetheless, following the rationale of Q-methodology presented in the methodology section, statistical significance is not prioritized a priori (Brown 1993; Watts and Stenner 2005). Additionally, regarding the threshold on the variance explained by each factor, Factor 1 and Factor 3 have eigenvalue > 1, whereas Factor 2 has eigenvalue = 0,994. In an attempt to provide a more detailed description of the perspectives on AI among Estonian civil servants, the Factor 2 eigenvalue score has been deemed sufficient to include this factor in the analysis. This is consistent with the preference for judgmental rotation in Q-methodology, which allows the researcher to explore alternative rotations of factors to discover insightful results (Brown 1993; Watts and Stenner 2005). In this research, and consistently with the methodology, the basis for judgmental rotation consisted of a combination of two elements. First, is theoretical knowledge of the matter. Second, are insights emerging from the answers of participants to the post-sort questions. From the theoretical

perspective, exploring the view of users of AI systems supporting decision-making promises to yield original results. Therefore, a factor rotation that would allow grasping the views of such participants in a factor was sought. In this sense, such participants are EE7, EE9, and EE12. From the perspective of insightful interviews, some framings of AI systems elicited the researcher's interest. For instance, EE8 stated clearly at the beginning of the interview, "*I am not an AI expert, but I trust technology and I trust people that work on technology and who work on new solutions and I believe that technology is a very helpful tool for society and the public sector as well and we have to look at different ways on how to use technology and at the same time how to be humans.*" Significantly, the idea of exploring the uses of technology and different ways of being humans was considered worthy of further exploration. EE5 said, "*If I have a grievance with the decision that has been taken, then I can look into it more thoroughly, but a lot of these systems that are already in place, they don't use AI, but they are meant to optimize workloads. I think these systems* (AI systems) *are less corruptible they are better in so many different ways. So optimization is meant to serve the process. But again, blind trust? No.* […] *In any process, person or system."* Significant here was the refusal of blind trust in any circumstances in parallel with a comparison between AI systems and other technologies in terms of complexity and lack of clarity of the processes. Given these insights, a factor rotation that would allow grasping the view of participants EE8 and EE5 in a factor was sought. Following these considerations, the most interesting result was found in a rotation of Factor 1 and 3 of +10°. Eliciting factors that would grasp such perspectives was possible only to a certain extent. Factor rotation, in fact, does not change the position of Q-sorts in relation to each other (Brown 1993; Watts and Stenner 2005). In other words, the relative difference or similarity between the Q-sorts stays unvaried when rotating the factors. What changes is the extent to which a factor accounts for a given Q-sort and, given the position of this Q-sort in relation to others, for the other Q-sorts. In practice, this means that, for instance, EE9 and EE12 load onto one factor while EE7 loads onto another factor. In the same way, EE12, EE9, and EE5 resulted loading onto the same factor. EE12 is the defining Q-sort of this factor, whereas EE5 loads onto that only to a limited extent due to its relative position against EE9 and EE12. In other words, EE9 and EE12 perspectives can be reduced to a common view on AI; this view is different from the one of EE7. EE5 perspective can be reduced to the same view as EE12 and EE9, albeit to a lower extent. A different rotation of factors would have grasped these perspectives differently, but it would not have changed the fact that the perspective of EE5 is somewhat similar to the one of EE12 and EE9. Using a different rotation of factors would have grasped different elements underpinning the participants' views on AI. Based on such elements, factors would have represented the same participants' perspectives but from a different angle. Nonetheless, the judgmental factor rotation allowed to account for

all the interesting perspectives highlighted before. Importantly, it was possible to include also Factor 2, although its eigenvalue is < 1. This was deemed useful for the research objective because EE8 loaded onto Factor 2. It should be re-stated here that this procedure is consistent with the use of Q-methodology and its purpose. This factor rotation resulted in two bipolar factors: Factor 2 and Factor 3. Bipolar factors are factors that are both positively and negatively correlated to various Q-sorts. This means that a bipolar factor is an expression of the viewpoint of Q-sorts that are positively correlated to that factor and of the opposite viewpoint, namely of those Q-sorts that are negatively correlated to that factor (Brown 1993; Watts and Stenner 2005). For instance, Factor 2 view is defined by the idea that the reliability of AI systems is fostered by the engagement of people affected by the system. Since Factor 2 is bipolar, both perspectives agree and disagree with such idea are grasped by this factor. In other words, Factor 2 represents views that, regardless of whether they agree or disagree with the idea that ethics of AI is fostered by the engagement of people impacted by the AI systems, are distinguished from other views by deeming this an important element. Splitting bipolar factors allows transforming the negative loadings into positive loadings, thereby creating two sub-factors loaded by the disagreeing and agreeing Q-sorts, respectively. Splitting bipolar factors is not a methodological requirement as, statistically, it does not change the percentage of variance in the data explained by one factor (Brown 1993). Similarly, on the conceptual level, it does not change the relationship between the main factor and the Q-sorts that load onto that (Brown 1993). Therefore, factor rotation resulted in 3 factors, of which Factor 2 and 3 were split into factor 2a and 2b, and 3a and 3b. Table 3 shows the factor loading of each participant with the three factors. Factor loading is the correlation between a given participant with a given factor (Brown 1993). In other words, it accounts for how much of the variance of a given participant's perspective can be explained by that factor. Values marked with an X correspond to the factor loadings of participants representing each factor. According to Q-methodology, for a participant to represent a factor, the statistical significance of the factor loading must be $p<0.05$, and the square of the correlation must be at least double half its common variance. However, consistently with the considerations on statistical significance and rich explanation of perspectives of Estonian civil servants on AI made before, Factor 2 has been assigned two representative Q-sorts. Because of this distinction between the statistical validity of Q-sorts representing Factor 1 and 3 on one side and Factor 2 on the other, Q-sorts that also fulfill the statistical validity criteria are marked with an *. Thus, as shown in Table 3, EE13 and EE8 are marked only with an X as they represent Factor 2, but they do not fulfill the statistical validity criteria. Table 3 shows how splitting bipolar factors does not impact the factor loading of the Q-sorts, aside from the relation sign. This is visible across Q-sorts and not only concerning the Q-sorts representing that factor. For instance, EE13 represents Factor 2b. EE13 factor

loading on Factor 2a is -0,5496, while its factor loading on Factor 2b is 0,5496. In the same way, EE13 factor loading on Factor 3a is -0,2506 while its factor loading on Factor 3b is 0,2506. In conceptual terms, two sub-factors now grasp the two poles of Factor 2. The first one is Factor2a, which expresses a view on AI systems that values the engagement of people impacted by the AI system in developing ethical AI. The second one is Factor 2b, which expresses a view that disagrees with the idea that ethical AI is enhanced by the engagement of people impacted by the AI system. Significantly, both Factor 2a and 2b are defined by this idea, albeit they are characterized by positive and negative feelings about it, respectively.

	FACTOR 1	FACTOR 2A	FACTOR 2B	FACTOR 3A	FACTOR 3B
EE12	0,7326* X	-0,0097	0,0097	0,0488	-0,0488
EE9	0,6714* X	-0,1387	0,1387	-0,2752	0,2752
EE10	0,5908* X	-0,3933	0,3933	-0,3125	0,3125
EE4	0,5541* X	-0,1443	0,1443	0,2625	-0,2625
EE2	0,5402* X	0,2278	-0,2278	-0,0724	0,0724
EE6	0,5327* X	0,2237	-0,2237	-0,0185	0,0185
EE5	0,4204* X	-0,0742	0,0742	-0,1815	0,1815
EE13	0,3379	-0,5496	0,5496 X	-0,2506	0,2506
EE8	0,2521	0,3398 X	-0,3398	0,1742	-0,1742
EE14	0,4573	0,2781	-0,2781	-0,5753*	0,5753* X
EE7	0,1655	-0,2821	0,2821	0,4095* X	-0,4095*
EE1	0,3132	0,2172	-0,2172	0,377* X	-0,377*
EE3	0,2557	0,2649	-0,2649	-0,3578*	0,3578* X
EE11	0,1132	-0,0135	0,0135	-0,321*	0,321* X

Author's own elaborate

Tab. 3 Factor Loadings (factor loadings with statistical significance $p<0.05$ are marked with * - representative participants for each factor are flagged with X)

Table 4 shows the factor scores correlation. Factors score correlation is the correlation matrix between the distinct factors extracted after the rotation (Brown 1993; Watts and Stenner 2005). In this case, statistically significant correlations are found between 1-2b and 1-3b. It is important to remember that as the factor structure accounts for different

underlying factors that explain the variance in the data, the lower the correlation between factors, the clearer the picture depicted by the factor structure is. In this sense, if the correlation between factors scores is low, those factors relate to different elements that can explain the variance in the data, hence the perspectives expressed among the Q-sorts. In other words, the correlation between factors expresses the similarity between them; hence it is an indicator of the similarity of the elements that could explain the viewpoints grasped by those factors and the viewpoints thereof. Table 3 shows that the factor scores correlation between Factor 1 and Factor 2b is 0,4689. Looking at the two views grasped by these factors, one could notice this similarity as both factors see the danger of AI systems increasing as the performance and reliability of the systems improve because this could make it harder for human operators to spot abnormalities. At the same time, both views believe that AI systems using self-learning improve the quality and accuracy of decision-making. Interestingly, these seemingly contrasting views within both factors could find a common explanation in both factors. Indeed, both Factor 1 and Factor 2b strongly disagree that AI systems with high-quality performance can be considered infallible. This brief account of Factor 1 and Factor 2b viewpoints is meant as an example intended to clarify the meaning of factor scores correlation in conceptual terms. A more thorough presentation of the factors' viewpoints will follow in this section.

	FACTOR 1	**FACTOR 2A**	**FACTOR 2B**	**FACTOR 3A**	**FACTOR 3B**
FACTOR 1	1	0,146	0,4869	0,2778	0,4014
FACTOR 2A	0,146	1	-0,1482	0,2446	0,1495
FACTOR 2B	0,4869	-0,1482	1	-0,1068	0,1565
FACTOR 3A	0,2778	0,2446	-0,1068	1	-0,0993
FACTOR 3B	0,4014	0,1495	0,1565	-0,0993	1

Author's own elaborate

Tab. 4 Factor Scores Correlation

The weight of Q-sorts loading on one factor is a metric that shows the relevance of each q-sort loading on a single factor for the definition of that factor. In other words, it expresses how much a q-sort defines the perspective identified by that factor. From a statistical standpoint, this metric shows how much a Q-sort defines the factor it loads onto compared to other Q-sorts assigned to that factor (Brown 1993). Beyond reporting the weight of a given Q-sort in defining the factor it loads onto, this metric is pivotal from a

conceptual perspective. In the interpretation of results, factor sorts weights provide orientation as to which participant's post-sort answers are more relevant to the factor related to that participant's perspective. Table 5 shows which Q-sort factor score has the highest weight on each factor. In Table 5, the Q-sort that weights the most on each factor is considered the defining Q-sort for that factor. Factor 2, after the split, retains one Q-sort per each sub-factor. Hence only one Q-sort loads onto Factor 2a; the same holds for Factor 2b. These Q-sorts are marked with an #. Factor 1 is mainly defined by EE12, factor 2a by EE8 (only Q-sort loading onto that), factor 2b by EE13 (only Q-sort loading onto that), factor 3a by EE7, and factor 3b by EE14.

FACTOR	DEFINING Q-SORT
1	EE12
2A	EE8#
2B	EE13#
3A	EE7
3B	EE14

Author's own elaborate

Tab. 5 Defining Q-sort per Factor (summary of factor sorts weight – Factors loaded by only one Q-sort are marked with #)

Another metric related to this is the intra-factor q-sorts correlation score, available in the data analysis report. This metric expresses the correlation between the q-sorts within each factor. In other words, it shows the similarity between the perspectives expressed by the Q-sorts that load onto a factor. This is important as it goes beyond the factor loadings of each Q-sort and indicates the homogeneity or heterogeneity of the perspectives common to a given factor. It is worth highlighting that the perspectives grasped by Factor 3a are highly heterogeneous as the intra-factor Q-sorts correlation score between EE7 and EE11 is 0,11.

4.2 Factor Interpretation

A holistic factor interpretation is made of two parts (Brown 1993). The first part consists of identifying consensus and distinguishing statements (Brown 1993). Consensus statements are those that participants see similarly, either in agreement or disagreement with the statements. Distinguishing statements are those that participants see in different ways. As the name suggests, the latter are the statements that distinguish the factors' viewpoints. The second part consists of an overview of the composite Q-sorts and the factors scores combined with the results of the post-sort questions. Since distinguishing statements are a crucial element of factors viewpoints, they are reported with the combined overview of the composite Q-sorts and the post-sort questions. In addition to

the report on the statistical analysis made so far, the overview of factors viewpoints combined with the answers to the post-sort questions that follow in subsection 4.2.2 suffices for the goal of this part of the dissertation. A detailed description of the results concerning the theory underpinning the research is presented in the discussion of results.

4.2.1 Consensus Statements

Table 5 reports consensus statements. Data analysis elicited two consensus statements with statistical non-significance equal to $p<0.005$. These are the statements that do not distinguish between any pair of factors. Namely, all the factors represent views that consider these statements neutral or lean towards agreement. Specifically, the opinion about statement n.30 is neutral across factors. In contrast, factors tend to agree with statement n.35. With regards to the statements that participants are primarily neutral about -i.e., score = 0- the most common reason given by the participants for classifying a statement as such is that this was not speaking to them. Possible reasons identified across participants are either lack of understanding about the aspect mentioned in the statement or lack of experience with it. For instance, EE13 explicitly refers to statement n. 30 when saying that it has not engaged with some aspects of AI systems provided in the Q-sample and hence scored these as 0. However, EE5 provides a different perspective on this statement, saying that it "…*implies that AI systems are open to sort of deterioration because the context is changing. But that's also true for everything. Everything has the same problem. […] Setting up a system that only has […] people accepting paper applications for whatever process, that also can deteriorate because the context is changing…"*. In other words, consensus about these statements indicates that –across factors- participants either feel they do not have an opinion about them or tend to agree with the point made in the statements. It is worth noting that EE13 also refers to statement n.35 in the example of statements that did not speak to it. This, for instance, provides an exception to the overall tendency to agree with statement n.35 across factors. This clarification on the exception helps to contextualize the relevance of consensus statements. More broadly, it shows that factor analysis inevitably draws commonalities across various perspectives, thereby reducing the richness of information the data collected in the research provides.

STATEMENT NUMBER	STATEMENT
30	Not only AI systems may change over time but also the context, this can make the performance deteriorate.
35	Within an AI system for decision-making, algorithms may be used as a convenient default for human decision-making, thereby reflecting wellknown mechanisms of satisficing behaviour.

Author's own elaborate

Tab. 6 Consensus Statements

Distinguishing statements are identified factor by factor. Indeed, the elicitation of a distinguishing statement is intrinsically related to one factor in that it distinguishes that factor from others. The statistical significance threshold for distinguishing statements has been set at $p<0.05$. Additionally, the Z-score of distinguishing statements is relevant to the interpretative task of Q-methodology. Given a statement, Z-score represents the deviation of a factor score of that statement from the factor score of other factors. Specifically, the bigger the Z-score -positive or negative- the more relevant that statement is for interpreting a factor viewpoint. Conceptually this means that a factor score with a big Z-score on a statement differs significantly -either in agreement or disagreement- from the other factor viewpoints on that statement. It is crucial to keep an eye on Z-scores while interpreting factors viewpoints as distinguishing statements are extrapolated based on a threshold set in their statistical significance -i.e., $p<0,05$. However, distinguishing statements with lower statistical significance -e.g., $p<0,1$- may have a more significant Z-score. This will be highlighted in the interpretation of factors viewpoints that follow.

4.2.2 Composite Q-Sorts Combined with Post-Sort Answers

This section examines the composite Q-sorts of each factor in combination with insights derived from participants' answers to the post-sort questions. Answers of participants whose Q-sort is defining a configuration -i.e., whose weight on the factor scores the highest among the Q-sorts representing such factor- are prioritized. Factor analysis with the centroid method emphasizes the configuration of ideas compared to traditional factor analysis (Brown 1993; Watts and Stenner 2005). Nonetheless, from a methodological perspective, the results of the data analysis are addressed as factors. Hence, the term factor is used in this section. However, as the primary goal of the research is eliciting configurations of ideas, the results shall be understood as such, consistently with the Q-Methodology. Additionally, said configurations are named after crucial characteristics that appear to be expressed to facilitate their understanding. The names are Aware Pragmatist, Human-Centred Innovator, Rational Executor, Heedful Explorer, and Critical Examinator.

The Aware Pragmatist

The following distinguishing statements identify Factor 1:

STATEMENT NUMBER	STATEMENT
36	Automation is most dangerous when it behaves in a consistent and reliable manner most of the time. In accordance, especially in the case of AI systems automating routinized tasks, human decision-makers at screen-level are likely to face difficulties in identifying abnormalities that warrant the use of their discretion.
3*	Development of technology and computation of data for and by AI are a cause of environmental and social issues.
25	AI systems using selflearning can lead to improved accuracy and quality of decision-making

Author's own elaborate

Tab. 7 Distinguishing Statements Factor 1 (Disagreement marked with *)

The preliminary consideration of distinguishing statements and the factor score indicates that the Factor 1 configuration's difference from other factors' configuration revolves around a disregard for the environmental and social concerns of AI and is somewhat concerned about the control that human operators can exert on the technology. However, using AI systems using self-learning can result in higher quality and accuracy of decision-making. Importantly, statement n. 36 is statistically significant at $p<0,05$, but its Z-score is 0,73. On the contrary, statement n. 3 is statistically significant at $p<0,1$, but its Z-score is -2,1. Similarly, statement n. 25 is statistically significant at $p<0,1$, but its Z-score is 2. Following a balance between statistical significance and richness of interpretation, statement n.36 clearly distinguishes Factor 1 from other factors. Nonetheless, statements n.3 and n.25 appear to be pivotal in interpreting Factor 1 configuration.

Seven out of the fourteen participants' views load on Factor 1, making this the most prominent factor. Factor 1 does not see technology, AI in particular, as a threat to humans. AI is not a competitor of humans at work (8: -3) nor a cause of environmental or social issues (3: -3). EE4 comments on this aspect, saying, "*I don't think that AI is a competitor of humans at work […], they are not against* (humans and AI*). I think that we should think about like […] a powerful tool which helps like in everyday life*." This is in accordance with the disagreement that technology's fundamental problem is the simplification of reality through numbers (2: -2). However, the possibility that AI is used as a convenient default for decision-making and humans fall into the mechanism of satisfying behavior by accepting everything that is suggested is not dismissed (35: 0). Indeed, this factor is distinct from the others by the idea that AI can become most dangerous when it is performing at its best (st. 36: +1). EE9 says, "*I think it's always important that a person checks what the AI has discovered. But I really do believe that AI can spot things that the human eye can't, or if they are so little that you just can't spot*

them, or sometimes just there are so many things that influence a person's decision-making. For example, if you have slept enough during the nighttime, …" Significantly, factor 1 viewpoint expresses a slightly positive attitude towards the possibility that AI is an asset for organizations to deliver on their mission (24: +1). One reason for this is the potential of AI systems based on self-learning to improve the accuracy and quality of decision-making (25: +3). Importantly, in some fields, these two characteristics are connected through a causal relationship, EE9 *"...again, in my field, if we are more accurate, then the quality also raises.*" There is reason to believe that this varies between organizations: EE10 on this aspect emphasized that "*In our organization, the amount of data is so big, so the need to use it and make conclusions are must*." Despite the positive connotation of technology that Factor 1 configuration depicts, AI is not seen as a panacea, and its use is undoubtedly not free from risks and imperfections (27: -3 and 7: -2). Concerns are mostly related to the computational power of AI systems, which makes it impossible for humans to oversee the correct functioning of the system in real-time (32: +3). Additionally, changes in the system and the context where it is used can deteriorate the performance, thereby limiting the benefits observed in the first place (30: +2). Pivotal for the challenges to be addressed and the risks to be managed is the overall approach to using AI that should ensure responsibility for using such systems. Significantly, the cornerstone of this approach lies in soft and hard regulatory approaches and organizational culture. An ethical code of conduct for developers is seen as a must to address responsibilities in AI systems (16: +3). Responsibility, in turn, is a crucial factor, EE9 *"...in the field of medicine there is a lot of responsibility and all the technology we are using has to support us and our decision-making and one life can depend on it.*" Nonetheless, policies that assign responsibility and ensure accountability should be implemented (18: +2). Similarly, purpose-oriented auditing for ethical AI has a slightly positive connotation in factor 1's view (21: +1). Indeed, Factor 1 configuration disagrees with relying solely on the qualifications of information officers (13: -1) or organizational leadership (15: -2) to ensure AI is used ethically. This is in accordance with disregarding the importance of informal relationships for data-sharing agreements (10: -2). EE4's opinion enriches the interpretation of the perception on this aspect in that it says, "*I think that those data sharing agreements should be formal. […] Yes, we could think that we will start with the informal relationship and informal sharing but in the end, you probably will have problems.*" Nonetheless, this is also influenced by the context under which the study is conducted; for instance, EE10 commented on this statement during the ordering process that its meaning was not fully clear as in Estonia data sharing within the public sector follows specific legislation. For factor 1, it is perhaps better if ethical AI is ingrained in the organizational culture (14: +1). In this sense, agencies should emphasize

the explainability of AI systems for those impacted by them (39: +2). Namely, it is essential that accuracy is explained and not only ensured (33: +2).

The following quotation may summarize the Aware Pragmatist's configuration of ideas:

In our organization, the amount of data is so big, so the need to use it and make conclusions are must. I think it's always important that a person checks what the AI has discovered. But I really do believe that AI can spot things that the human eye can't or sometimes just there are so many things that influence a person's decision-making.

The Human-Centred Innovator

The following distinguishing statements identify Factor 2a:

STATEMENT NUMBER	STATEMENT
12	Reliability of AI system comes from the engagement of people affected by the system
15	Agencies need a leader responsible for keeping a focus squarely on ethical AI practices and fostering the coordination to make it happen.
37*	AI systems make decisions often using complex mathematical equations. It is not possible to guarantee that the decisions these systems make are always ethical, especially when they operate on new data and possibly in changing environments once they are deployed.

Author's own elaborate

Tab. 8 Distinguishing Statements Factor 2a

The preliminary consideration of distinguishing statements and the factor scores indicate that Factor 2a configuration's difference from other factors' configurations is centered on the idea that the implementation of ethics in AI systems is not hindered by the complexity of AI systems nor by new datasets or changing environment. Factor2a viewpoint associates ethical AI systems with the engagement of people impacted by the AI system and the existence of a leader in the organization that makes sure that practices related to ethical AI are implemented. All the statements are statistically significant at $p<0,05$, with statements n. 12 and n.15 that score even lower. Regarding the Z-score, statement n.15 scores 1,78 making this statement central to the interpretation of Factor 2a viewpoint.

Only one participant's view is represented by Factor 2a, namely that of EE8. However, it is important to note that Factor 2 is bipolar and includes the views of two participants that are grasped by Factor 2a and 2b, respectively. Factor 2a view highlights the importance of the human component in technological development and, specifically, in the use of AI technology. This prominence is also expressed by the disagreement with the view of AI as a competitor of humans at work (8: -3). Said human component is embodied in different elements of the use of AI systems. The cornerstone for an ethical approach to

AI appears to be the role of organizational leadership (15: +3). EE8 on this "*Responsibility of leaders? I think that is quite important.*" Similarly, the engagement of the people affected by the system plays a significant role in ensuring the reliability of AI (12: +3). Significantly, the view expressed by Factor 2a seems to take a step further on the engagement of stakeholders: EE8, "*I trust technology and actually I trust people who work on technology and on new solutions...*". Nonetheless, the stakes are high in technological development, and smart technology might not yet be fit-for-purpose (1: +1). However, Factor 2a view considers a combination of factors in this sense: EE8, "*…technology is a very helpful tool for society and for the public sector as well and we have to look at different ways on how to use technology and at the same time, how to be a human*." It is perhaps for this reason that in the view of Factor 2a the use of AI should follow clear regulations, for instance, policies that enforce a framework of responsibility and accountability should be put in place (18: +3). The connection here appears to be societal trust in technology, EE8 *"...every member of society, uh, who uses technology has to trust it*." Finding successful use cases may be challenging in practice because not everyone has a sufficient level of data literacy or because the indicators of success have been set without considering the stakeholders (11: +2). Eventually, "*It's very difficult to change. It's very human but very often people take it as a concern, but it should be taken as a challenge*." (EE8). Moreover, even when AI's performance satisfies the success criteria, this can change over time due to changes in the context (30: +2). Relevant to successful development appears to be the ethical use of AI, EE8 *"...ethical issues are also very, very important.*" The relationship between the elements of trust, change, and ethics appears to be expressed in the agreement of Factor 2a's view with the importance of reporting and advice channels in the creation of trustworthy and inclusive AI (17: +2). The collaborative approach is also valued beyond the organizational boundaries between sectors (20: +2). In this context, Factor 2a's view leaves the door open for informal relationships (10: 0). Overall, openness and collaboration should translate into prioritizing the creation of algorithms that can be explained to those they impact (39: +2). Despite the challenges and risks highlighted, Factor 2a's view deems it feasible for humans to take on the challenges and succeed, such as those related to information-sharing agreements (19: -3). From the perspective of Factor 2a, the ethical use of AI seems to be a matter of finding the right way to proceed, guaranteeing that ethical principles are respected (37: -2). This is perhaps best exemplified by Factor 2a's view on large-scale AI systems. Large-scale deployment of AI is challenging (6: +1) but manageable, as the disagreement with the concern of the large-scale impact of inaccuracies seems to suggest (28: -3). In this sense, if accuracy does not outweigh explainability, neither is the latter perceived as intrinsically fundamental, if not in establishing a baseline of trust as reported above (29: -2 and 33: -1). Overall, the economic value of AI should be kept in mind when considering

the concerns (7: +1). EE8 on this, "…*if we concern a lot, we just lose possible economical value*." In this context, and about organizations delivering on their mission, AI is considered as a tool, and it is the human intelligence that is an asset for organizations (24: -2). However, solidifying the relationship between AI and humans expressed in the interview, EE8 explains that AI is perhaps an asset for human intelligence.

The following quotation may summarize the Human-Centred's configuration of ideas:

Technology is a very helpful tool for society and for the public sector as well and we have to look at different ways on how to use technology and at the same time, how to be a human. I trust technology and actually I trust people who work on technology and on new solutions… Responsibility of leaders? I think that is quite important.

The Rational Executor

The following distinguishing statements identify Factor 2b:

STATEMENT NUMBER	STATEMENT
36	Automation is most dangerous when it behaves in a consistent and reliable manner most of the time. In accordance, especially in the case of AI systems automating routinized tasks, human decision-makers at screen-level are likely to face difficulties in identifying abnormalities that warrant the use of their discretion.
26	It is difficult to follow-up on AI predictions for decision-making because they can generate accurate but counterintuitive insights due to the large number of variables and data they use. They go against the grain of traditional heuristics.
38	Since AI systems making decisions can consist of multiple components possibly developed by various organizations and based on data across different sources, it can be impossible to identify responsibilities for errors. This can hinder accountability.
20*	Cross-sectoral cooperation is essential for trustworthy AI systems for decision-making.
12*	Reliability of AI system comes from the engagement of people affected by the system.

Author's own elaborate

Tab. 9 Distinguishing Statements Factor 2b (Disagreement marked with *)

The preliminary consideration of distinguishing statements and the factor score indicates that Factor 2b viewpoint's difference from other factors' viewpoints is focused on concerns over the feasibility of acting upon AI predictions and disagreement with the idea that engaging people impacted by AI systems or cross-sectoral cooperation make the system reliable and trustworthy. Moreover, Factor 2b viewpoint's difference from other factors' viewpoints revolves around concerns on the control of human operators over well-performing systems and the accountability of AI systems decision-making. All the distinguishing statements exceed the statistical significance threshold, with statements n. 12, n.20, and n.26 scores even more significant. Statement n.12, n.26, and n.36 are the most relevant for interpreting Factor 2b viewpoint as they score ±1,78 on Z-score.

Only one participant's view is represented by Factor 2b, namely that of EE13. However, it is essential to note that Factor 2 is bipolar and includes the views of two participants that are grasped by Factor 2a and 2b, respectively. The core point of connection between Factor 2a and 2b appears to be the view of AI and people in organizations, specifically with regards to the delivery of the mission of an organization. Namely, both poles of this factor believe that AI is a tool that helps people, but it is the latter that delivers the mission (24: -2). In the words of EE13, "*I think it is important to distinguish that AI helps people and organizations to deliver the mission but is not an asset itself. I like to think about it as a tool.*". At its roots, this shared understanding seems to be underpinned by the belief that technology, AI, in this case, is no magic and should not be used solely for the sake of using it: EE13 "*Also, AI is not always a solution, and should not be done just to have AI.*" In this sense Factor 2b represents a more realistic view of the risks, challenges, and limitations of AI. This perspective is perhaps crystallized in the acknowledgment of EE13 that using AI for decision-making is difficult as it goes against "*gut feelings*" (26: +3). EE13 says, "*I have seen that if the output is not confirming the traditional heuristics, people tend to think it's data problem, not the problem of traditional heuristics. It's very complicated to prove otherwise, […] people (in our case, consultants) are rather relying on their gut (heuristics) than predictions from the AI system.*" A focal point is the explainability of accuracy in AI systems (33: +1), however, "…*it's hard to explain what algorithm the model exactly uses and how the output is calculated*." (EE13). The issue is the computational power of AI systems that make it impossible to oversee it in real-time and identify mistakes (32: +2), EE13 "…*even if you can explain a lot of what the AI system is doing, there will be always something that is very hard to control or check if done correctly. This also decreases the reliability of the AI system and makes it harder to really make decisions based on the outputs.*" Factor 2b view highlights the complexity of algorithms and how this results in the difficulty of spotting problems and mistakes (37: +3). EE13 on this "…*in my experience it's very hard to detect and find the mistakes/problems in the AI system, because the data goes through such a long and complex process.*" The complexity of AI systems and their value chain also creates problems with accountability and responsibility (38: +2). This is not an aspect to be disregarded in the configuration resembled by the Factor 2b view, as the scalability of AI decision-making systems can potentially affect millions of stakeholders (28: +2). Besides this, AI can be dangerous even when performing consistently and reliably, arguably, this is even more worrisome as it can prompt satisfying mechanisms in the human operator (36: +3). Nonetheless, AI decision-making systems can generally be trusted (4: +2). EE13 says, "*You have to trust the algorithm...*" but also acknowledges that *"...it might be hard, especially if the user does not understand fully what is happening.*" (EE13). Eventually, Factor 2b view highlights the problem of wrong expectations and misunderstandings

deriving from the complexity of AI. In the words of EE13 "…*people want and expect AI and other technology to be very reliable and flawless but that's never the case. And with AI it is important to remember (especially if it's predicting probability of something happening like our tool) that nothing is happening with 100% probability and that's hard for people to understand*." This is indicative of Factor 2b's expression of disagreement with the idea that the reliability of AI systems comes from engaging the people affected by the system and that the system can be assumed to be infallible when its performance increases (12: -3 and 27: -3). The perspective on the challenges and limitations of AI systems generated by the complexity of the technology are reflected in other elements of the Factor 2b view. It seems that using an ethical code of conduct for developers, reporting or advice channels and cross-sectoral cooperation can do little to increase the transparency and trustworthiness of the system (16: -1 and 17: -1 and 20: -1).

The following quotation may summarize the Rational Executor's configuration of ideas:

You have to trust the algorithm but it might be hard, especially if the user does not understand fully what is happening. People want and expect AI and other technology to be very reliable and flawless but that's never the case. And with AI it is important to remember that nothing is happening with 100% probability and that's hard for people to understand. I have seen that if the output is not confirming the traditional heuristics, people tend to think it's data problem, not the problem of traditional heuristics. It's very complicated to prove otherwise.

The Heedful Explorer

The following distinguishing statements identify Factor 3a:

STATEMENT NUMBER	STATEMENT
39*	Within an AI system for decision-making, agencies should emphasize creation of algorithms that are transparent and can be explained to people who are being impacted by them.
1*	Smart technology is still not always ready to address some concerns of societal benefits such as the ones agreed on in the SDGs.
11*	There are people who have no clue where to begin. Either because they have no data literacy, or because no one has documented what this is supposed to mean; or because someone else decided this is the indicator of success, and they don't necessarily believe that.
9*	New professions in the public sector linked to AI will have to be tackled thoroughly and other "intelligences" or capabilities will be required from civil servants.

Author's own elaborate

Tab. 10 Distinguishing Statements Factor 3a (Disagreement marked with *)

The preliminary consideration of distinguishing statements and the factor score indicates that Factor 3a viewpoint's difference from other factors' viewpoints is centered on the

idea that there is no need to focus on civil servants' capabilities despite there are cases of low data literacy or loosely defined objectives with regards to AI success factors. Additionally, Factor 3a viewpoint's distinction revolves around a rejection of the idea that technology is not ready to address societal issues and explainable algorithms should be prioritized. All the distinguishing statements exceed the statistical significance threshold, with statements n.39 and n.9 being even more significant. Statement n.9 appears as central for the interpretation of Factor 3a viewpoint.

Only two participants' views are represented by Factor 3a, namely that of EE7 and EE1. However, it is essential to note that Factor 3 is bipolar as well and includes the views of five participants that are grasped by Factor 3a and 3b, respectively. Factor 3a views' core is that AI systems for decision-making can be trusted (4: +3). The main reason appears to be the capabilities of such systems, especially those AI systems using self-learning that can increase the accuracy and quality of decision-making (25: +2). In this sense, accuracy is considered a vital criterion against which AI systems should be evaluated (29: +1), whereas the explainability of algorithms to those impacted by the AI systems should not be a priority for organizations (39: -1). Factor 3a seems rather pragmatic regarding how AI operates and what this implies for society. This can be seen in the slight agreement with considering that smart technology that simplifies reality into numbers is a problem (2: +1). However, in the words of EE7, "*Pointing out some details outweighs all the concerns that you might have with artificial intelligence*." Importantly, this is not a matter of the economic value of AI (7: -2), at least not today: EE7 "*Maybe also after 10-15 years it will become more relevant if we have less and less doctors, nurses. Then if you have something like, for example a patient can use their own like decision maker tool*." Instead, an important benefit of AI systems for decision-making today appears to be related to the possibility opened by AI systems to elicit human bias in decision-making (5: +2). More broadly, Factor 3a views are characterized by a disagreement with the idea that smart technology cannot solve issues related to societal benefit (1: -2). The same distinction between the present and future of AI systems can be seen regarding the idea of AI as an asset for organizations to deliver their mission (24: -3). Factor 3a expresses a realistic but optimistic view on this. EE7 says, "*Maybe in 10-15 years, but not today. […] I think it takes time and before family doctors will see the value of this […] at the moment it's like background information*." Reasons could be found in the lack of data literacy and competencies among civil servants (11: -2). However, this is not the case for everyone: Factor 3a view does not see AI insights as complex to follow up on (26: -2). Moreover, according to Factor 3a, the lack of competencies is not an impediment but a temporary constraint, and there is no need to address it thoroughly (9: -3). It appears to be a matter of turnover, EE7 "*This problem will be solved by itself.*" Beyond benefits and constraints to the use of AI, Factor 3a view values an ethical approach to AI. Foremostly, policies

play a significant role in ensuring responsibility and accountability (18: +3). This is clearly expressed by EE1: "*AI needs laws and clarity of responsibilities, otherwise there will be a problem later on who is responsible for AI activity.*" Responsibility is central to Factor 3a's view, and an ethical code of conduct for developers is considered a must-have to address it (16: +3). Regarding this aspect, cross-sectoral cooperation is also valued in Factor 3a perspectives (20: +2). Importantly, this has implications for the further development of AI use in the public sector: EE7 says, "*I hope that one day we'll have this really cross sectional that you will have lots of different information available. The decision-making tool we have is only focusing on the diseases and the guidelines we have, but I would like it if it were attached to the social system as well, all the like for example, […] habits, patients have they can self-report to the system...*" Indeed, Factor 3a's view leaves the door open for informal relationships to set up data-sharing agreements (10: 0) and, in general, does not see information-sharing agreements as a constraint that can lead to issues that cannot be solved (19: -3).

The following quotation may summarize the Heedful Explorer's configuration of ideas:

And then you think 'oh really good that you have this: like someone who's reminding you this small but important things'. And if the system uses self learning it's also important: it can correlate it and next time it will give me even better advice and the quality of decision-making making will improve definitely. I would delete the idea that AI is an asset. Maybe in 10-15 years, but not today. […] at the moment it's like background information.

The Critical Examinator

The following distinguishing statements identify Factor 3b:

STATEMENT NUMBER	STATEMENT
13	It is important to know that the information officers that are working with the AI system are qualified people.
33	Explanations on accuracy are important in AI systems for decision-making.
12	Reliability of AI system comes from the engagement of people affected by the system.
32*	AI systems for decision-making can process more information at a higher speed than even the best trained humans are capable of. There is no way in which the human operator can check in real-time that the computer is following its rules correctly.
23*	Technical standards are the main source to promote and safeguard ethical values in AI.
31*	Stability of an AI system making decisions is particularly important in the public sector, where many external factors affect decision-making.
38*	Since AI systems making decisions can consist of multiple components possibly developed by various organizations and based on data across different sources, it can be impossible to identify responsibilities for errors. This can hinder accountability.

Author's own elaborate

Tab. 11 Distinguishing Statements Factor 3b (Disagreement marked with *)

The preliminary consideration of distinguishing statements and the factor score indicates that Factor 3b viewpoint's difference from other factors' viewpoints is a solid agreement with the importance of explainability of AI and the qualifications of professionals working with AI systems combined with a firm refusal of the idea that the complexity of AI systems hinders accountability. This is consistent with the distinguishing disagreement on technical standards as the primary source to ensure that AI is ethical. This, in turn, is consistent with disagreement on using technical standards as the main source to ensure ethics. Additionally, Factor 3b viewpoint's distinction lies in the disagreement with the idea that humans cannot control AI in real-time and that the stability of AI systems making decisions is particularly relevant to the public sector. All of the distinguishing statements exceed the statistical significance threshold, with statement n.33 being significant at $p<0,01$. Statement n.13 appears central to the interpretation of Factor 3a viewpoint with a Z-score of 2,13. Similarly, statements n.31, n.38, and n.33 constitute crucial aspects of Factor 3b viewpoint as they score -1,7, -1,79, and 1,96 on the Z-score, respectively.

Only three participants' views are represented by Factor 3b, namely that of EE14, EE11, and EE3. However, it is important to note that Factor 3 is bipolar as well and includes the views of five participants that are grasped by Factor 3a and 3b, respectively. Factor 3b views distinguish themselves from Factor 3a views in that they are not convinced that AI systems for decision-making can be trusted (4: -1). EE14 says about this, "*I can't say like I trust completely...*" but specifies *"...I think I would take it like with open mind and I would try it definitely*." The reason could be found again in the words of EE14 *"...It learns, and it can learn, I don't know, falsely*." This opinion is reflected in the disagreement with the idea that these systems can be deemed infallible with improving

the quality of automated decision-making of AI systems (27: -3). More specifically, using AI systems does not necessarily improve the quality of decision-making, for instance, highlighting biases that could not be spotted in human decision-making (5: -2). Factor 3b does not seem to assign any particular importance to the scalability of AI systems in such considerations as they disagree that large-scale deployment typical of AI systems is challenging per se (6: -2). The problem highlighted by Factor 3b views appears to be the scarce possibility of ensuring an ethical approach due to the complexity of AI systems and the environment they most likely operate in (37: +3 and 30: +2). For instance, EE3 clarifies this aspect with an example when it says, "…*as technology improves and then develops so quickly and there are so many pathways and guidelines actually supporting this* (the use of AI systems for decision-making), *which are also important things to... that should be actually considered so whether we can ensure that all, all new guidelines are already involved in this system.*" Nonetheless, Factor 3b perspectives disregard the importance of stability of AI systems, at least as a peculiarity of their use in the public sector (31: -3). In the context of the other factor scores, this specification seems to be relevant as it is reasonable to think that stability is valued disregarding the context, which is why this factor score is –3. Importantly, the reason for the limitations seen in AI systems by Factor 3b views is not a fundamental one of technology (2: -2). The comment of EE3 on this appears insightful *"... I would say that it supports rather than simplifies. The reality is that there are patients who do not agree or will not visit a family physician and they have chronicle conditions, but it's all their choice, whether they will come to the visit or whether they will use healthcare services. So, this developed technology really improves or supports everyday activities at the primary care level*." Following this, the possibility that AI's economic value outweighs its use concerns is left open (7: 0). The point seems to be the responsibility for AI systems decisions tout court. Indeed, in the view of Factor 3b perspectives, responsibility is not hindered by the multitude of actors potentially involved in its development (38: -3). Nonetheless, responsibility must be addressed through an ethical code of conduct for developers (16: +2). More broadly, the importance of the fact that information officers working with AI systems are qualified professionals is highlighted by Factor 3b views (13: +3). This is a far better option than using technical standards as the main source to promote and safeguard ethical values in AI (23: -2). Perhaps, Factor 3b perspectives' positive view on the role of people in ensuring ethical AI could also be noticed in the disagreement with the idea that AI computational power makes it impossible for the human operator to check it for mistakes against the rules of the system (32: -1). With an educated guess, it could be argued that this is hinted at by EE14 in that it says, *"...and if it remembers the fault and it goes from there, then we have a trouble*." considering that only unfixed mistakes by the system are problematic. However, the fact that human operators might fall into the trap of satisfying

behavior is not excluded (35: 0). Interestingly, within this context, there is some distrust towards the need for purpose-oriented auditing for ethical AI (21: -1). Another relevant aspect of Factor 3b's perspective is the explanation on accuracy and of the algorithms to those impacted by AI systems (33: +3 and 39: +2). Aside from the emphasis organizations should put on the development of explainable AI (39: +2), the role of cross-sectoral cooperation is important for trustworthy AI systems (20: +2). Nonetheless, Factor 3b views value the role of policy to establish clear responsibility and accountability to a certain extent (18: +1). This, perhaps, is seen as a safeguard against the fear of ending up in information-sharing agreements that create problems that cannot be solved (19: +2).

The following quotation may summarize the Critical Examinator's configuration of ideas:

I can't say like I trust completely. I think I would take it like with open mind and I would try it definitely. It learns, and it can learn, I don't know, falsely... and if it remembers the fault and it goes from there, then we have a trouble.

In order to answer the research questions, the following section discusses these configurations of ideas on trust in AI systems used in the public sector in light of the literature on trust in technology and the use of AI in the public sector.

5 Discussion of Results

The main objective of this research is to provide an account of extant perspectives on Artificial Intelligence among civil servants. Notably, a subset of technologies under the umbrella of Artificial Intelligence was considered in this research. Said subset was defined on three dimensions derived from Van Noordt (2022). First, the technique used by the technology was limited to Machine Learning due to its significant autonomy from human intervention. Second, the capabilities of the technology were limited to perceiving, learning, and making decisions due to their resemblance to human behavior, somewhat unique to AI. Third, the application of the technology was limited to recommendation systems due to its relevance in the workflow, as human operators have to act upon the recommendation of a technological tool. It must be noted that the last dimension is not unique to Artificial Intelligence technology. Nonetheless, the first two dimensions constitute unique elements for said application, in contrast to other recommendation systems that rely on different techniques and capabilities. This subset of technologies and applications is consistent with the AI taxonomy proposed by the AI Watch, whose ultimate goal is to identify the range of technologies worth further investigation as belonging to the family of Artificial Intelligence technologies (Samoili et al. 2020). Given the focus of this research on civil servants, a survey among Belgian civil servants was used as a reference to define the subset of AI technologies (Van Noordt 2022). The population under study is Estonian civil servants. This choice was made because of the advancement of some AI projects within the Estonian public sector. This thesis identifies five configurations of ideas on trust in Artificial Intelligence systems used in the public sector among Estonian civil servants, namely: the Aware Pragmatist, the Human-Centred Innovator, the Rational Executor, Heedful Explorer, and the Critical Examinator. The configurations of ideas consist of ideal-types representations of how Estonian civil servants interconnect elements of trust in Artificial Intelligence. Additionally, the five ideal-types shed light on how Estonian civil servants configure the elements of trust in Artificial Intelligence systems used in the public sector with the pre- and post-adoption phases of Artificial Intelligence in the public sector.

The findings of this thesis contribute to the extant literature on Artificial Intelligence in the public sector in two ways. First, the findings shed light on civil servants' subjectivity about the use of Artificial Intelligence systems in the public sector. These findings add to the knowledge on the matter in that civil servants' perceptions on AI are overlooked in research (Ahn and Chen 2021). Second, the findings account for the perspective of trusters in the truster-trustee relationship. These findings add to the knowledge on the matter in that extant research focuses on the characteristics of trustworthy AI, disregarding the trusters' subjective perceptions (Zhu et al. 2021). Importantly, this

constitutes a significant gap in that there is a potential mismatch between the trustworthiness of AI and the civil servants' perception of such trustworthiness (Zhu et al. 2021). Additionally, the findings of this research provide insights into the pre- and post-adoption phases of Artificial Intelligence in the public sector. They do so in that they shed light on their perspective on the organizational, institutional, and environmental influencing factors.

The discussion of results will be structured as follows. First, the five configurations of ideas are discussed regarding trust in Artificial Intelligence systems used in the public sector. Second, factors influencing AI's pre- and post-adoption phases in the public sector are examined and put in relation to the five configurations. Finally, limitations and implications for further research will be discussed at the end of this section.

5.1 Estonian Civil Servants' Configurations of Ideas on Trust in Artificial Intelligence Systems Used in the Public Sector

The appropriate contextualization of the findings of this research on civil servants' trust in AI is the discovery of specific configurations of ideas regarding trust in AI systems - hence, estimations of trustworthiness (Zhu et al. 2021)- among Estonian civil servants. In light of this contextualization, general comments and comparisons on such configurations of ideas can be put forward across factors viewpoints. The underpinning of trust in technology is a general disposition toward technology (McKnight et al. 2011). This consists of the belief that by using technology, one can achieve better outcomes than by not making use of it (McKnight et al. 2011). This is consistent with the conceptualization of cultural and individual context put forward by Lee and See (2004) as the underpinnings of a favourable view of technology in general. On this dimension, it can be observed that The Aware Pragmatist, The Human-Centred Innovator, and The Rational Executor viewpoint express a favourable disposition towards technology (see subsection 2.3.1), albeit showing some doubts. On the contrary, The Heedful Explorer and The Critical Examinator represent a less favourable disposition towards technology (see subsection 2.3.1). Provided that all participants belong to the same cultural context, it appears that different individual contexts may account for such differences (see subsection 2.3.1). However, conclusions cannot be drawn about what kind of individual context (see subsection 2.3.1) leads to such differences because a pattern cannot be identified among participants. Institutional-based trust in technology consists of the belief that the use of a specific technology is likely to be successful because an individual already feels comfortable using a generic type of that technology and because the specific technology is embedded in a system of support, safety, and guarantee mechanisms (McKnight et al. 2011). Zhu et al.'s (2021) clarification that such elements are to be found in the processes

and people of an organization appears relevant for the matter at hand. This is consistent with what Lee and See (2004) identify as organizational context. Indeed, McKnight et al.'s (2011) conceptualization of institutional-based trust appears to be somewhat limited to organizational structures and excludes the human component. In this dimension, the findings appear consistent with the participants' experiences using smart technology and their account of support systems and processes across factors viewpoints (see subsection 2.3.1). For instance, The Aware Pragmatist viewpoint considers the existence of policies and regulations on the use of technology in the public sector as necessary to ensure trust, such as the ones on data collection and data sharing in Estonia. On the contrary, The Human-Centred Innovator viewpoint mainly focuses on the role of people to steer innovation and technological development in a direction that benefits society, such as organizational leaders. Significantly, The Human-Centred Innovator grasps the view of the Head of Innovation Department in an Estonian public sector organization. Differently, the Rational Executor disregards the elements of institutional-based trust (see subsection 2.3.1) in Artificial Intelligence. This configuration primarily focuses on the elements of trust in a specific Artificial Intelligence system, suggesting that trust is heavily influenced by the specific characteristics of each AI system, according to the Rational Executor's configuration. Conversely, the Heedful Explorer's configuration points to safeguards on the institutional-based trust dimension (see subsection 2.3.1) as a central element of trust in Artificial Intelligence. It seems that this configuration highlights said dimension as it seeks safeguards from potential threats caused by using Artificial Intelligence. Interestingly, The Critical Examinator perspective on institutional-based trust (see subsection 2.3.1) shows some ambiguity as to the influencing factor of the experience with a generic type of technology of which AI systems are a specific instance pointed out by McKnight et al. (2011). Importantly, this configuration grasps the perspective of two participants who express doubts regarding the success of technology in their field -i.e., education.

Trusting beliefs in a specific technology, as conceptualized by McKnight et al. (2011), are compared and combined with the conceptualization of trust in automation proposed by Lee and See (2004) to conceptualize trust in Artificial Intelligence. Importantly, this research does not attempt to measure the level of trust of civil servants in AI nor the magnitude of its impact on the willingness to explore a technology (McKnight et al. 2011; Thatcher et al. 2011). The findings on this dimension provide insights into how civil servants combine elements of trust in the specific technology and institutional-based trust (see subsection 2.3.1) in their configurations of ideas on trust in AI systems used in the public sector. For instance, The Aware Pragmatist viewpoint values elements such as functionality in trusting AI systems (see subsection 2.3.1). This appears to be coupled with a focus on a system of support, safety, and guarantee mechanisms that ensures the

reliability of AI (see subsection 2.3.1). Differently, The Human-Centred Innovator seems to prioritize people in the institutional-based trust and combine this with a minor focus on the functionality of AI, suggesting that trust is given to the people behind the system on the institutional level of trust (see subsection 2.3.1). On the opposite side, the Rational Executor configuration of trust in AI seems to derive from the functionality of AI systems for public sector organizations, whereas (the lack of) information on the process hinders it (see subsection 2.3.1). In this case, institutional-based trust does not seem to balance these elements (see subsection 2.3.1). The Heedful Explorer's configuration of trust in AI focuses on the reliability of AI systems (see subsection 2.3.1), consistently with its sought of safety. Specifically, the reliability of AI systems should be coupled with safety mechanisms on the institutional-based dimension of trust (see subsection 2.3.1). It is worth noting that whereas the Aware Pragmatist and Human-Centred Innovator's configurations appear to seek a balance between elements of trust in the specific AI systems and the institutional-based dimension of trust (see subsection 2.3.1), the Heedful Explorer values an alignment between the elements of these two dimensions. The Critical Examinator values elements of reliability and the understanding of the internal processes of AI systems (see subsection 2.3.1). It appears that this configuration comes from a sense of responsibility that the Critical Examinator assigns to itself as a (potential) user of AI systems. This configuration expresses an ambiguous perception of the institutional-based dimension of trust in AI (see subsection 2.3.1). This ambiguity appears to lead to uncertainty regarding the effectiveness of the institutional-based elements of trust (see subsection 2.3.1), which in turn leads to focusing on diving deeper into the characteristics of AI systems and valuing them as reasons to decide whether to trust AI systems or not.

Provided this comparison of the five configurations, this section continues with a more detailed examination of each of the configurations by highlighting more specific elements of the three dimensions of trust in Artificial Intelligence and referring to the specific experiences of participants in the selected AI projects in the Estonian government.

Aware Pragmatist

The aware pragmatist disregards potential societal issues deriving from using technology while refraining from considering it useful upfront. More specifically, the aware pragmatist seems to formulate its trust in AI around a combination of safety mechanisms and technology functionality.

The aware pragmatist is characterized by a favourable view of technology in general (McKnight et al. 2011). This favourable view implies a tendency to believe that technology usually leads to better outcomes (McKnight et al. 2011). The aware pragmatist viewpoint maintains a neutral position towards considering the readiness of smart

technologies to address societal issues. This configuration doubts whether smart technology can always be viewed positively upfront. The individual and cultural context are deemed influential on dispositional trust in technology (Lee and See 2004; McKnight et al. 2011). On the one hand, these results are consistent with the cultural context of this study. Arguably, the case of Estonia is unique in its use of technology in public services, providing a cultural background for such disposition. On the other hand, the individual background of civil servants loading on the Aware Pragmatist is partially consistent with such disposition. Four out of seven participants do not have a background in IT. Nonetheless, their roles in their respective projects or organizations put them in close contact with IT professionals. For instance, one is a user trainer of an AI system for decision-making in social affairs, while another is responsible for technology use in an Estonian high school. More interesting is the case of a project manager in the healthcare sector, who explicitly mentioned that this was their first experience managing projects related to digital public services. Elements of institutional-based trust (see subsection: 2.3.1) are not standing out for the Aware Pragmatist. However, it is worth noting that the Aware Pragmatist somewhat disregards the human component of the institutional dimension (see subsection 2.3.1). In Lee and See (2004), this is associated with the organizational context. Conclusions on this matter cannot be drawn with the data available for this research because no element points at an organizational context that may discourage trust in the human component (Lee and See 2004; Zhu et al. 2021). In fact, only one participant loading on the Aware Pragmatist admittedly mentioned a lack of shared understanding of the processes to ensure the robustness and safety of AI systems as a constraint to developing an AI project in the field of education. It is worth mentioning that also another participant loading on the Aware Pragmatist belongs to the public sector organization impacted by that. Nonetheless, the literature's clearest characterization of institutional-based trust in technology refers to a system of support, safety, and guarantee mechanisms (McKnight et al. 2011). In this sense, the Aware Pragmatist considers elements such as policies, code of conduct for developers, and cross-sectoral cooperation sources of trust in AI. Specifically, these aspects of institutional-based trust (see subsection 2.3.1) are viewed as ensuring accountability. This appears consistent with the comments from participants loading on this configuration. Indeed, some of them referenced data collection and sharing laws in Estonia as an example of existing support from policies. Others, instead, explicitly stated how only such mechanisms could ensure trust in AI and technology in general. The elements of the dimension of trust in the specific technology (see subsection 2.3.1) stands out for the Aware Pragmatist. Specifically, Factor 1 viewpoint shows a positive attitude toward the functionality of AI systems for the functions carried out by participants (see subsection 2.3.1). However, concerns over the performance and reliability of AI systems are expressed by the Aware

Pragmatist (see subsection 2.3.1). The so-called black-box appears to be the reason of such concerns whereas the utility of insights provided by AI systems may be the origin of positive views towards their functionality (see subsection 2.3.1). Positive feelings towards explainability and transparency of AI systems shown by the Aware Pragmatist are consistent with this interpretation. These findings are consistent with the experience of the two participants weighting the most on the composition of this configuration. These participants have direct experience of the AI systems used in their fields -i.e., social affairs and healthcare- being a user trainer and a field expert engaged in the development of the tool and now user. Both participants commented positively on the usefulness of the AI tools used in their fields. At the same time, they bewared of risks regarding unconditional trust and overreliance. The relationship between institutional-based trust and trust in the specific technology appears of particular interest to the Aware Pragmatist (see subsection 2.3.1). This could be interpreted as trust in the institutional dimension -significantly regarding agreements and policies and not people- safeguarding against risks posed by the specific technology (see subsection 2.3.1). It is to say that those mechanisms to assign responsibility and accountability ensure that if the individual cannot effectively oversee the technology, there is still a system accounting for it.

Human-Centred Innovator

The Human-Centred Innovator considers whether smart technology is ready to address societal issues as an important factor determining its trust in AI. It appears that the trust in AI of the Human-Centred Innovator relies on the role of leaders and professionals in ensuring that the technology is harnessed in favor of development and human wellbeing.

The Human-Centred Innovator shows a mainly favourable disposition toward technology (McKnight et al. 2011). Nonetheless, its view is more tuned compared to the Aware Pragmatist. The disposition to trust technology expressed by the Human-Centred Innovator reflects on whether AI for decision-making can be trusted upfront. More importantly, it accepts the idea that smart technology is not always ready to address the most pressing issues for society. This result is partially consistent with the cultural context of this study but perhaps more so with the individual context of the participant loading on this configuration (Lee and See 2004; McKnight et al. 2011). If the use of technology in the Estonian public sector is considered background for a favourable disposition towards technology, then the Human-Centred Innovator may have a less optimistic perspective on the general usefulness of technology. However, this can be seen differently when considering the individual context of the participant loading on this configuration. With the role of Head of Innovation Department, EE8 is perhaps more prone to thorough considerations regarding technology. This interpretation is consistent with the

participant's comments on their experience. Indeed, a generally positive view towards technology is put in the perspective of the complexity of innovation and the responsibility of leaders and innovators in using technology to benefit society. The Human-Centred Innovator shows trust in the institutional dimension of trust in technology (see subsection 2.3.1). Regarding the human component (see subsection 2.3.1), this appears to be consistent with the organizational context, highlighted by Lee and See (2004) as influencing trust in technology. In this sense, the role of leaders is highlighted as well as the engagement of people impacted by the system. This appears to be consistent with the importance assigned to innovators and people working with technology expressed by EE9. Significantly, EE9 explicit trust in the people in such roles. Institutional-based trust in technology is also expressed through a positive view of support, safety, and guarantee mechanisms (McKnight et al. 2011). Specifically, cross-sectoral cooperation and advice channels are seen in a positive light. With an educated guess, this appears to be the result of the participant's experience with the generic type of technology of which AI systems are a type (McKnight et al. 2011). EE9 referred to some projects where smart technology is successfully used in education. However, this is true only to a limited extent as the characterization of AI and smart technology lacks solid fundaments, thereby making this argument rely solely on the account made by the Human-Centred Innovator. The Human-Centred Innovator also presents concerns on the elements of institutional-based trust in technology (see subsection 2.3.1). These are mainly centered on aspects of data governance. The Human-Centred Innovator deems information-sharing agreements as a potential source of issues. Additionally, the misalignment of goals across levels of the organization may be a problem for the Human-Centred Innovator. Interestingly, this might come from some individuals' lack of data literacy, contrasting with the stark comments on trusting people working with technology. The Human-Centred Innovator shows several concerns about trust in the specific technology, hence specific AI systems (see subsection 2.3.1). Trust in the reliability of AI systems is limited due to the changes in the systems themselves and the context of deployment (see subsection 2.3.1). Additionally, the accuracy of AI systems does not outweigh the value of explainability. It appears that the Human-Centred Innovator highlights concerns over the reliability and functionality of AI systems, whereas it disregards potential issues associated with the internal processes of AI systems (see subsection 2.3.1). Indeed, the Human-Centred Innovator does not consider complex algorithms a source of concern. These results may be interpreted in light of the institutional-based dimension of trust (see subsection 2.3.1) and the role of the participant loading on this configuration. Significantly, the position as Head of Innovation Department qualifies EE9 as a public manager. Consistently, EE9 reported its experience in leading projects rather than using AI systems in its daily workflow. This might account for a focus on the institutional dimension of trust in

technology (see subsection 2.3.1). Additionally, this may also explain why the Human-Centred Innovator does not see AI systems as assets for public sector organizations -with specific reference to delivering on their mission. The Human-Centred Innovator appears to focus on the system put in place to use AI systems as the ultimate determinant of AI usefulness, rather than seeing said systems by themselves as pivotal for public services.

Rationalist Executor

The Rationalist Executor represents a demystified view of technology in general for its fundamental simplification of reality into numbers. The formulation of trust in AI made by the Rationalist Executor combines a positive view of the functionality of the technology with careful consideration of the complexity of the process.

The Rationalist Executor shows a generally favourable disposition toward technology (McKnight et al. 2011). Nonetheless, its view is more tuned compared to the Aware Pragmatist. The Rationalist Executor accepts the idea that smart technology is not always ready to address the most pressing issues for society. This can be put in perspective with the doubts of the Rationalist Executor on whether technology suffers from a fundamental problem, hence the simplification of reality into numbers. As for the Human-Centred Innovator, the Rationalist Executor appears to express less disposition to trust technology than what could be expected given the cultural context (Lee and See 2004; McKnight et al. 2011). Similar to the Human-Centred Innovator, the individual context of the single participant loading on this factor may provide further insights (Lee and See 2004; McKnight et al. 2011). EE13 is an analyst in a public organization in the field of social affairs. Regarding the AI-related project, EE13 is also in charge of change management. It results that the Rationalist Executor seems to express a technical and utterly demystified view of smart technology and technology in general. Interestingly, this shows a partially contrasting view compared to participants loading on the Aware Pragmatist with an IT background. However, it is worth noting that the nature of the role of these civil servants differs quite radically as EE13 primary role is that of a data analyst. With an educated guess, this may account for a more demystified view of the potential of data-driven decision-making. The Rationalist Executor disregards the institutional dimension of trust (see subsection 2.3.1) in that this does not play a major role in what this viewpoint deems relevant on whether to trust AI systems. Regarding the generic type of technology of which AI systems for decision-support are a specific instance (McKnight et al. 2011), the Rationalist Executor does not show particular concern. This is consistent with the experience of EE13 with such technology as it is part of its daily work (McKnight et al. 2011). However, the human component of institutional trust in technology is viewed with concern by the Rationalist Executor (see subsection 2.3.1). Conclusions on this dimension

cannot be drawn as no data points to an unfavorable organizational context (Lee and See 2004). Additionally, the fact that the Rationalist Executor somewhat disregards this dimension altogether suggests that this may not be relevant for this viewpoint. Of more importance are the elements of the dimension of trust in the specific technology, hence AI (see subsection 2.3.1). The Rationalist Executor expresses concerns about multiple elements of the dimension of trust in AI systems (see subsection 2.3.1). However, this appears to be compensated by a positive view regarding the functionality (see subsection 2.3.1). Indeed, the Rationalist Executor ultimately sees positively the potential benefits of AI for the functions carried out by public sector organizations, such as decision-making. However, this is not witnessed in reality. It is to say that for the Rationalist Executor the complexity of the internal process of AI systems that makes the use of its insights almost impossible seems to outweigh completely the potential benefits (see subsection 2.3.1). This is consistent with the agreement shown by the Rationalist Executor with the idea that accuracy does not outweighs explainability. Additionally, it appears that the so-called black-box creates a fundamental problems as even when AI performs well the human operators cannot exert full control hence limit its use.

Heedful Explorer

The characterization of trust of the Heedful Explorer builds on a balance between potential benefits and threats deriving from the use of technology in general. The formulation of trust in AI of the Heedful Explorer values a system of safety mechanisms, with the contribution of professionals, as much as the reliability and functionality of the technology itself.

The Heedful Explorer shows a disposition to trust technology somewhat characterized by ambiguity (McKnight et al. 2011). On the one hand, the rejection of concerns about the readiness of technology to address societal issues is coupled with a favorable view of the trustworthiness of AI for decision-making. On the other hand, Heedful Explorer doubts the implications of the simplification of reality intrinsic in technology and the environmental and social issues possibly arising from the development of AI. It derives that the disposition toward technology (McKnight et al. 2011) of the Heedful Explorer results as the least favourable perspective encountered in this study. This appears to be somewhat surprising given the abovementioned considerations in the Estonian cultural context. However, the individual context of the participants loading on The Heedful Explorer may provide some insights (Lee and See 2004; McKnight et al. 2011). EE1 participated in developing a pilot for an AI system for decision-support in the field of emergency services in the information manager role. Nonetheless, EE1 was assigned this role because of its expertise in the field and not because of an IT background. Similarly,

as a field expert, EE7 was involved in developing an AI system for decision support in healthcare. Beyond these considerations, both participants also expressed mildly positive views about technology in their comments on their experience. Based on this, this demystified, somewhat ambiguous disposition towards technology can be reconducted to their acknowledgment of the potential benefits of technology and its potential threats combined. In this sense, EE7 specifically mentioned the peculiarity of its field -i.e., healthcare- as a reason to be careful with the use of technology without this hampering innovation. The Heedful Explorer appears to have a favourable background of trust in technology rooted in the elements of the institutional dimension of trust (see subsection 2.3.1). It appears that the participants loading on the Heedful Explorer have a positive experience with the use of smart technology, highlighted by McKnight et al. (2011) as influencing institution-based trust. Indeed, they do not see it as a competitor of humans nor as a trigger for new "intelligence" to be sought amongst civil servants. Significantly, the Heedful Explorer deems AI systems as a solution to spot and eliminate biases in human decision-making. The Heedful Explorer shows institutional-based trust also in the human component and the mechanisms of support and safeguards (see subsection 2.3.1). Favorable opinions are expressed regarding the importance of qualifications of professionals working with technology and cross-sectoral cooperation to develop AI systems. Notably, The Heedful Explorer shows a strongly favorable opinion on the role of policies and regulations in assigning responsibility for AI systems operations. Such findings are consistent with the experience reported by the participants loading on the Heedful Explorer. Both of them commented positively on their experience with smart technology, suggesting that they comfortably use the generic type of technology of which AI is a specific instance (McKnight et al. 2011). Additionally, both projects benefitted from inter-organizational cooperation, suggesting that a positive experience with supporting systems and IT professionals constitutes a background for institutional-based trust (see subsection 2.3.1). The Heedful Explorer values the elements of the dimension of trust in the specific technology (see subsection 2.3.1). AI systems for decision-support are seen as functional regarding public sector organizations' functions such as decision-making (see subsection 2.3.1). Additionally, the complexity of the internal processes of AI systems appears to be disregarded by The Heedful Explorer (see subsection 2.3.1). However, the Heedful Explorer strongly rejects the idea that AI is an asset for public sector organizations. The reliability of AI systems appears as another central element of trust in AI systems for decision-support for the Heedful Explorer (see subsection 2.3.1). Such findings are consistent with the views expressed by participants loading on the Heedful Explorer, as one of them is a satisfied user of AI for decision-support in healthcare, and the other one categorized the pilot project of an AI system for decision-support in the field of emergency services as successful. It is to say that both participants

reported a positive experience with the use of the technology and see its use as benefitting their work.

Critical Examinator

The Critical Examinator disregards the technology in general as an element of trust in AI. The formulation of trust in AI of the Critical Examinator shows uncertainty over the value of safety mechanisms and focuses more on the possibility of understanding the processes internal to the technology.

The disposition to trust technology in general of the Critical Examinator appears low (McKnight et al. 2011). Despite disregarding the potential issue of simplifying reality intrinsic in technology, the Critical Examinator does not show a favourable disposition towards technology. Significantly, technology is not seen as always ready to address societal concerns. Regarding the cultural context (Lee and See; McKnight et al. 2011), the considerations made for the other factors apply, thereby making the findings on the Critical Examinator the least consistent with the cultural context of this study. However, these are perhaps the most consistent with the individual context of participants loading on the factor (Lee and See 2004; McKnight et al. 2011). Two out of the three participants loading on the Critical Examinator are teachers in Estonian high schools -i.e., EE11 and EE14. Both of them expressed strong doubts about the usefulness of technology in their field. This might be due to the lack of empathy and the ability to understand the context. The third participant loading on the Critical Examinator -i.e., EE3- was involved in a project to develop an AI system for decision-support in healthcare as a field expert. Interestingly, EE3 explicitly said that even if technology simplifies reality, this is not a problem as this technology feature ultimately supports healthcare practitioners in their daily work. Therefore, EE3 appears to express a somewhat shifted perspective on the usefulness of technology in its field, albeit adopting the same rationale as EE11 and EE14 on technology in general. The Critical Examinator viewpoint is somewhat ambiguous on the elements of the dimension of institutional-based trust in technology (see subsection 2.3.1). The uncertainty on aspects such as the need for new professions in the public sector and the economic value of AI is coupled with disagreement on the idea that AI can debias human decision-making. However, the large-scale deployment typical of AI is not a cause of concern. This is somewhat consistent with the reported use of such technologies (see subsection 2.3.1) by participants loading on the Critical Examinator. On the one hand, EE14 and EE11 mentioned they used certain tools for education that may qualify as smart technology, thereby being the generic technology that AI is a specific instance of (McKnight et al. 2011). On the other hand, EE3 was involved in developing an AI system for decision-support in healthcare since the early phases and is now responsible for the

adoption of that tool by general practitioners. Albeit not using the AI system itself, EE3's position appears to result in a perceived comfort in using such tools, which is highlighted in the literature as an influencing factor for institutional-based trust (McKnight et al. 2011). Therefore, uncertainty on whether using AI systems for decision-support is likely to be successful or not appears consistent with the participants' background in that they have only a limited experience with the generic type of technology (see subsection 2.3.1). Similarly, the Critical Examinator appears uncertain about the elements of institutional-based trust related to mechanisms of safeguard and support systems and the human component (see subsection 2.3.1). On the one hand, cross-sectoral cooperation and ethical guidelines for developers are seen in a positive light. On the other hand, information sharing agreements are a reason for concern, and ingraining ethical AI in the organizational culture is not considered a valid option to ensure the ethical use of AI systems. These findings are consistent with participants' perception of such elements based on their experience. For instance, EE11 stated that the achievement of learning outcomes is not granted by using new tools such as AI systems for decision support. In saying so, EE11 referred to the fact that it is a process of trial and error, thereby suggesting that there is no support system in place (McKnight et al. 2011). Additionally, it mentioned that teachers remain responsible for the achievement of learning objectives, thereby making the use of new tools somewhat risky for their performance evaluation by the Ministry of Education. Interestingly, EE3 is also part of the support team for using the AI system for decision support in healthcare. EE3 mentioned that there had been some issues on the technical side in terms of integration of the AI system with existing tools used by general practitioners. The reason is resources for the integration of such systems in the workflow. Therefore it appears that EE3 experience is consistent with the findings as trust in the support systems is somewhat limited, albeit existent, due to shortcomings resulting from lack of resources (McKnight et al. 2011). The Critical Examinator viewpoint values transparency and explainability of the internal processes of AI systems as sources of trust in the specific technology (see subsection 2.3.1). This is also visible in the concerns expressed in the Critical Examinator viewpoint regarding the complexity of AI systems. Additionally, for the Critical Examinator, reliability is important in trusting an AI system (see subsection 2.3.1). This is combined with disregard toward the possibility that AI systems help the participants that load on the Critical Examinator in their daily work. These findings are consistent with the teachers' experience on how helpful AI can be in their field. However, it appears to be less consistent with the view of the field expert in healthcare. It is worth recalling here its perspective on the fact that the AI system for decision-support in primary care is deeply beneficial for general practitioners in that it improves their results and it is cost-effective.

5.2 Estonian Civil Servants' Configurations of Ideas on Trust in Artificial Intelligence Systems About Pre- and Post-Adoption Phases of Artificial Intelligence in the Public Sector

The concept of trust in technology is strictly related to technology adoption and post-adoption phases (see subsection 2.3.1). For instance, the perceived usefulness of technology is grasped by the element of functionality in McKnight et al. (2011) and performance in Lee and See (2004), as it is present in technology acceptance models (see: TAM). Findings on civil servants' configurations of ideas on AI systems used in the public sector shed light on how said individuals see challenges in the pre- and post-adoption phases of AI in the public sector. In practical terms, for instance, whether a configuration of ideas highlights the role of leaders or policies to ensure responsibility for AI systems and the ethical use of AI ultimately provides insights into how Estonian civil servants' configurations of ideas on trust in AI relate to elements of pre- and post-adoption phases of AI in the public sector.

5.2.1 Artificial Intelligence Pre-Adoption in the Estonian Public Sector

Literature finds the perspective of innovation diffusion to be a fruitful approach to understanding Artificial Intelligence adoption (Van Noordt and Misuraca 2020). Significantly, isomorphism -i.e., a tendency to alignment- is used as a lens to read the diffusion of innovation (Van Noordt and Misuraca 2020). Intergovernmental cooperation that occurs at the level of international organizations is a manifestation of such processes. For instance, the AI system for primary care in the field of healthcare object of this study is an example of such processes of diffusion of innovation. The project started after the proposal of the World Health Organization, as EE2 reported. EE2 was involved in the project as IT manager to ensure the successful development of the algorithm to improve primary care in Estonia. Specifically, the objective was to support general practitioners in spotting high-risk individuals and tailoring their interventions to their needs. Interestingly, EE2 reports that the cooperation was sought due to the advancements of the Estonian digital infrastructure, namely the availability of data in the public sector - specifically in healthcare. This is consistent with the antecedents of AI adoption found in the literature (see subsection 2.2.1). However, the first version of the algorithm was not suited to the Estonian digital infrastructure, and integration between the information systems was sought, according to EE2. Specifically, EE2 was involved in the project to ensure that the second version of the algorithm would be tailored to the digital infrastructure. It is worth noting that EE2 reported that the second version of the algorithm was developed on the bases of close collaboration between the national government and the World Health Organization. On the contrary, EE2 points to loose collaboration as the

cause of the unsuccessful development of the first version of the algorithm. The experience of EE2 appears to be consistent with The Aware Pragmatist suggesting the existence of the relationship between what public managers focus on and the context they find themselves in (Alsharani et al. 2021). The Aware Pragmatist configuration of trust positively values policies and agreements in developing AI systems (see subsection 5.1). According to the findings of Van Noordt and Misuraca (2020), inter-organizational data-sharing within the public sector and between the public and private sectors is an important environmental antecedent. The environmental push for inter-organizational cooperation seems to be coupled with institutional challenges (see subsection 2.2.1). Campion et al. (2020) highlight how data sharing is a reason for concern for governmental organizations. The literature points to the sensitiveness of data collected and used by governments as a reason for concern (Campion et al. 2020; Pencheva et al. 2020). Consistent with these findings is the Estonian Data Protection Agency's concern regarding the use of AI in education, as reported by EE6. Consistent with the experience of EE6, in the Aware Pragmatist's viewpoint, policies and regulations on data sharing should support the development of AI systems. In this case, it is worth noting that EE6 points to a tailored approach to policymaking data-sharing policies that account for and are tailored to new approaches to developing AI systems, such as open source. Organizational antecedents consist of structural and cultural factors (see subsection 2.2.1). Campion et al. (2020) observe that a lack of understanding of data needs is one of the key challenges to inter-organizational cooperation on AI adoption projects. The results of this research appear partially consistent with the knowledge on organizational antecedents (see subsection 2.2.1) in that the AI adoption project in healthcare seems to suggest a relationship between understanding data needs and inter-organizational cooperation. However, according to EE2's account of the AI adoption project in healthcare, the loose collaboration created a lack of understanding of data needs. Campion et al. (2020) find inter-organizational trust to influence the effectiveness of data sharing. Nonetheless, conclusions on this cannot be drawn as the lack of understanding of data needs appear to be the result of loose collaboration, but no insights on inter-organizational trust can be gathered from the interview with EE2. Changing processes and working routines at the organizational level impacts AI adoption (Campion et al. 2020; Penchaeva et al. 2020). Criado et al. (2020) find that the co-development of AI systems positively influences the adoption and use of such technology by first-line bureaucrats. The views expressed by the factor viewpoints appear consistent with these findings. The AI system for primary healthcare was co-developed with the support of general practitioners. Participants in this project with such roles load on The Aware Pragmatist and The Heedful Explorer. Both configurations express a favorable view of the AI system's functionality for the civil servants' daily work. Management support is another relevant antecedent, together with organizational

resources (Sun and Medaglia 2019; Van Noordt and Misuraca 2020). The Critical Examinator viewpoint shows uncertainty on institutional-based trust, particularly regarding support mechanisms (see subsection 2.3.1). EE3 describes its role in the AI project in healthcare as supporting the development of the functionalities of the AI system. EE3 is currently leading the project and is responsible for ensuring the successful adoption and use by general practitioners. Commenting on its experience, EE3 points to the lack of resources as a reason for shortcomings in the system's effective update. Consistent with the literature, it appears that EE3 experience contributes to an ambiguous view on institutional-based trust in that the existence of management support -namely, EE3 itself- is coupled with a lack of resources (see subsections 2.2.1 and 2.3.1). AI technologies may be adopted with different goals (Veale and Brass 2019). Framing AI systems in some ways rather than others is likely to influence the result of the adoption (see subsection 2.2.1). For instance, defining an AI system as a decision support system instead of a decision-making system is likely to help the acceptance of civil servants, hence its adoption (Criado et al. 2020). This is also related to indirect environmental factors in that conceiving AI as a well-suited solution to external triggers facilitates its adoption (Van Noordt and Misuraca 2020). On the contrary, Sun and Medaglia (2019) identify unrealistic expectations and misunderstandings of AI as challenges to its adoption. Regarding this, the healthcare AI system was defined as a decision-support system by the participants across factors viewpoints and regardless of their roles. In this light, the experience of general practitioners and their comments on the functionality of the AI system for their daily work appears to be consistent with findings on certain framings constituting a facilitator for AI adoption (Criado et al. 2020). Nonetheless, this appears not to be the case with the AI system in emergency services. The latter is framed as a decision-support system by EE1, but the project is stuck after a successful piloting phase. However, EE1 is not a user, and other factors may be more relevant in this project. Additionally, it seems that conceiving said AI system as a well-suited solution to address the problem of primary healthcare for high-risk individuals facilitated its adoption (Van Noordt and Misuraca 2020). Specifically, AI systems are seen as a solution due to their computational capability compared to humans. The experiences of these participants are translated into the factor viewpoints in that The Aware Pragmatist and The Heedful Explorer view AI systems functionality positively and as an important element for trusting AI (see subsection 5.1). In the case of The Aware Pragmatist viewpoint, considering AI as a well-suited solution for external triggers (Van Noordt and Misuraca 2020) appears to be consistent also with the experience of EE12 (see subsection 5.1). EE12 works in a public sector organization in the field of social affairs and expresses the need for using AI due to the massive amount of data to analyze to solve the wicked issue of unemployment.

5.2.2 Artificial Intelligence Post-Adoption in the Estonian Public Sector

Literature on the impact of AI in the public sector highlights the relevance of public value creation by the public sector -measured by performance, openness, and inclusion- (Van Noordt and Misuraca 2020). Scholars use this perspective to frame the research on the impact of AI in public sector organizations (Van Noordt and Misuraca 2020). In this sense, the opinion of civil servants on the impact of AI on public value creation is somewhat grasped by the element of functionality in the concept of trust in a specific technology (see subsection 2.3.1). Indeed, functionality grasps whether an individual deems "a technology to have the capacity or capability to complete a required task" (McKnight et al. 2011, p. 9). In this sense, it can be argued that whether civil servants find AI useful for their tasks sheds light on whether they see it as useful for the functions carried out by their teams and organizations. Therefore, it can be argued that this suggests whether they perceive that AI has a positive impact on public value creation (see subsection 2.2.2), albeit partially. The Aware Pragmatist and The Heedful Explorer viewpoints grasp general practitioners' views of public value creation (see subsection 2.2.2) in these terms. Specifically, both configurations positively view AI's functionality in civil servants' work (see subsection 5.1). This can be observed in the perspective of EE9, who says that AI helps general practitioners keep track and account for many small details in diagnosing patients. A reflection on The Human-Centred Innovator may provide additional insights on this aspect, specifically regarding the connection between adoption and post-adoption phases. The Human-Centred Innovator suggests a relative view on the functionality of Artificial Intelligence for civil servants (see subsection 5.1). It is to say that The Human-Centred Innovator appears to establish a relationship between the functionality of AI systems and the use that civil servants -significantly, under the guidance of leaders and innovators- make of it (see subsection 5.1). It appears that The Human-Centred Innovator does not evaluate the functionality of AI systems based on whether this is useful to complete required tasks but rather on whether said technology can be used to improve the creation of value from the public sector (see subsection 5.1). Specifically, EE8 -the participant loading on The Human-Centred Innovator- argues that creating value from the use of Artificial Intelligence in the public sector demands experimentation and *"...to look at different ways on how to use technology and at the same time, how to be a human."* (EE8). Thus, this configuration seems to suggest that achieving the successful implementation of AI in the public sector is a matter of establishing use-cases that create public value for citizens. This passage between the adoption and post-adoption phases centered on public value creation seems to align with the public value creation perspective found in the literature (Van Noordt and Misuraca 2020). Importantly, this may suggest that civil servants value this perspective when deciding whether to move past the adoption phase of AI. On the institutional level,

supportive policies and regulations positively affect AI's successful implementation and use (see subsection 2.2.2). For instance, well-suited data sharing agreements are identified as influencing the implementation phase other than the adoption (see subsection 2.2.2). This appears to be particularly relevant, albeit in opposite directions, in the case of EE12 and EE6. EE12 commented on the role of data-sharing regulation in Estonia as a facilitator for successfully implementing AI in the context of policies and regulations to assign responsibility. On the contrary, EE6 views policies and regulations on data as a determinant factor, but its current fashion is constraining. Importantly, EE6 is involved in an AI project in the field of education that appears to be stuck between adoption and implementation. This could suggest that it is specifically at the implementation phase that data-sharing agreements and policies are determinant (see subsection 2.2.2). Both participants are loading on The Aware Pragmatist, which expresses a positive view towards policies and regulation as sources of trust in AI (see subsection 5.1). Factors on the organizational level specific to the post-adoption phases (see subsection 2.2.2) cannot be highlighted from the results of this study. Campion et al. (2020) find that a lack of awareness on AI potential is a prominent factor in the organizational skills in this phase, whereas the lack of leadership skills is more relevant in the adoption phase. Whereas the Human-Centred Innovator is consistent with such findings in the adoption phase (see subsection 5.2.1), there are no solid insights regarding the impact of awareness on AI potential (Campion et al. 2020) on the post-adoption phases among configurations. The aspect of civil servants' skills (Campion et al. 2020) is highlighted only by participants in the field of healthcare. This is consistent with the relevance of such aspect in the post-adoption phases (Campion et al. 2020), as this AI project is already implemented. Specifically, EE7 refers to the lack of awareness of the usefulness of AI (see subsection 2.2.2) among general practitioners as the root cause for their refusal to use the AI tool for decision-support for primary care. However, the concept of trust in Artificial Intelligence (see subsection 2.3.1) used to frame the configurations refers to the skills of civil servants with a focus on other elements than the awareness of AI potential (see subsection 2.2.2). Therefore, conclusions on this matter cannot be drawn. The Heedful Explorer viewpoint -loaded by EE7- expresses a positive view of the human component of institutional-based trust (see subsection 5.1). Importantly, it must be noted that this element grasps the belief that there is a support system of safety and guarantee mechanisms that enhance trust in the technology (McKnight et al. 2011). Public managers generally have autonomy in implementing AI in public sector organizations, and civil servants directly interact with such technologies in their daily work (Veale and Brass 2019). Therefore, individual factors' influence is more important in the post-adoption phases (see subsection 2.2.2). The attitudes of individual public servants towards the specific AI system are found to have a greater impact in post-adoption phases as this influences the daily use of such

technology (Ahn and Chen 2021). This thesis's main objective was to explore civil servants' subjectivity on AI, specifically trust in AI. Regarding the literature on individual factors influencing the post-adoption phases specific to AI, this thesis contributes to extant knowledge in that it sheds light on existing viewpoints -i.e., preferences and expectations (Zhu et al. 2021)- among civil servants in Estonia. It is worth noting that only the AI projects in healthcare and social affairs have passed the adoption phase, thereby qualifying as projects in the post-adoption phases. Interestingly, configurations of ideas accomunate civil servants' perspectives across projects and sectors. This factors' structure suggests that in the context of this study, it cannot be argued that the use of AI in daily work is a key determinant of the configuration of ideas an individual adopts on such technology. However, the opposite cannot be concluded either. This study relies on Q-methodology, which provides a snapshot of the views expressed by individuals on a given topic (see subsection 3.1). Therefore, the possibility of changing ideas over time cannot be excluded based on the results of this study. It is to say that the findings of this study suggest that, to date, the views of Estonian civil servants are not determined by whether they use or not AI systems for decision-support in their daily work. Nonetheless, no scientific evidence supports the claim that this cannot happen in the future. Importantly, the AI projects that moved to the implementation phase were piloted no longer than five years ago. The lack of longitudinality constitutes a limitation of this research that might be addressed by conducting further research in the same setting. This and other limitations and directions for future research are discussed in the following section.

5.3 Limitations and Future Research

This research is affected by several limitations. These are presented in this section together with directions for future research. The latter is framed to solve some of the limitations encountered in this study and draw from this thesis to continue accumulating knowledge on the topic. Regarding the limitations, the Q-sample developed based on the concourse suffers from certain shortcomings deriving from its development. The concourse is meant to represent the broad spectrum of ideas on a given topic (see subsection 3.1.1). It is important that the statements of the Q-sample are selected and reworded so that they do not express facts but opinions (see subsection 3.1.1). The Q-sample for this thesis was developed by the researcher accordingly. However, some of the statements resulted resembling facts more than opinions. For instance, statement n. 28, "AI systems making decisions can be deployed at scale, possibly affecting millions of stakeholders. Any inaccuracies can translate into a large number of mistakes." Additionally, some statements could be interpreted as including multiple opinions. For instance, EE6 expressed this concern regarding statement n.28. This may also be true for

statement n.11 “There are people who have no clue where to begin. Either because they have no data literacy, or because no one has documented what this is supposed to mean; or because someone else decided this is the indicator of success, and they don’t necessarily believe that.” Whereas the statement provides possible causes for civil servants' disorientation on AI projects, participants may agree or disagree with only some of them and find it difficult to rank them. Additionally, according to some participants, the concourse resulted in a Q-sample including complex statements. For instance, EE5 asked to complete it after the interview as it deemed the statements difficult to understand. Similarly, EE2 commented on the statement that *“I’d have to think about them for a long time.”* The root cause for this limitation is the development of the concourse based mostly on sources including experts' opinions. Due to a lack of resources, the concourse was developed on the basis of academic literature, conferences, and publications on the topic gathered through the researcher's network. Importantly, the latter source of information for developing the concourse was included to add more hands-on and less sophisticated statements. Nonetheless, as shown, this objective was only partially achieved. Lastly, on the concourse and Q-sample, some of the statements associated with one dimension of the concept of trust in technology are not fully aligned with it. Provided that there is a lack of agreement on the definition of Artificial Intelligence and Artificial Intelligence technologies (see subsection 2.1), some statements are not fully aligned to the dimension of institutional-based trust (see subsection 2.3.1). Specifically, this applies to the statements meant to grasp a participant's comfort in using the generic type of technology that AI systems for decision-making are an instance of (McKnight et al. 2011). Regarding the Q-Sorting, respecting the fixed normal distribution resulted somewhat challenging for participants. This limitation was foreseen, and the research methodology section provided justification for its use (see subsection 3.2). However, it must be reported here that one Q-Sort was excluded from the data analysis because the participant explicitly said that respecting the fixed normal distribution was not possible. Regarding the methodology, also the P-sample suffers from limitations. Specifically, the P-sample appears to be too little heterogeneous. Sampling for Q-methodology is best done using purposeful sampling techniques (see subsection 3.2). In this research, the strategy consisted in contacting a first participant for each AI project addressed and following a snowballing approach afterward. The first participants were asked to provide colleagues' contact information that matched some characteristics that would have made the sample heterogeneous. Nonetheless, this was not always possible due to work overload or participation in other studies. In particular, the P-sample is not particularly heterogenous in that it includes eleven women out of fourteen participants. With regards to the field of public services that participants work in, there is a noticeable unbalance. Indeed, only one participant accounted for the use of AI in emergency services and two participants for social affairs.

This is against five participants related to the healthcare sector and six to education. Regarding the validity of the findings, there is one major limitation. Factors 2 and 3 account for few participants regardless of the factors split. Specifically, Factor 2 accounts for only two participants and Factor 3 for five participants. Such factors were kept because they accounted for at least one of the participants whose role in its organization or answers to the post-sort questions were deemed worthy of further exploration. Additionally, weighting the impact of this limitation on the findings, it should be considered that said threshold is more of a rule of thumb than a solid statistically relevant rule (see subsection 3.3). Nonetheless, if the statistical threshold of eigenvalue > 1 is applied to determine which factors are to be kept, then Factor 2 does not exceed the threshold as its eigenvalue equals 0,99. However, Factor 3 exceeds the threshold. Moreover, it is important to consider that The Heedful Explorer accounts for only two participants and that the intra-factor Q-sorts correlation score is only 0,11. It is to say that The Heedful Explorer grasps heterogeneous perspectives. While not being a limitation per se, combined with the low number of participants loading onto said factor, this constitutes a noteworthy shortcoming in its validity. Further research can be tailored to address some of the said limitations. Additionally, research can build on this study and further explore the socio-technical aspect of Artificial Intelligence in the public sector (see subsection 2.2.2). A brief discussion on both directions of further research follows here below. As noted above, some participants found it challenging to complete the Q-Sort. Two reasons identified during the interviews are the complexity of some of the statements of the Q-Sample and the constraints posed by the fixed normal distribution used for the Q-Sort. The statements' complexity can be reduced by adopting a different approach to constructing the concourse (see subsection 3.1.2). Therefore, further research seeking to use the Q-methodology may solve such limitations by developing the concourse based on interviews with a sample of the population under study (see subsection 3.1.1). Nonetheless, it should be noted that the Q-methodology aims to elicit original configurations of ideas, requiring purposeful sampling (see subsection 3.2). This requirement, combined with the complexity of the topic at hand, may still result in a concourse that includes statements that are difficult to understand for some of the participants. Further research may also use the concourse developed for this research and circumnavigate this limitation by developing a different P-sample. For instance, a P-sample including only civil servants involved in AI projects already implemented may solve the problem of little understanding of some statements. Nonetheless, this is plausible only to some extent as purposeful sampling requires the inclusion of participants with different backgrounds (see subsection 3.2) -hence, for instance, knowledge of common practices or topics in the field of AI systems development. Regarding the problem encountered with the fixed normal distribution adopted in this study, further research can address it by using a free distribution. As

presented in the research methodology section, this brings benefits and other limitations (see subsection 3.1.2). Alternatively, further research could use a wider scale for a fixed normal distribution. This will allow for more margin for participants to agree or disagree with the statements on both sides of the distribution. Another limitation of this study, intrinsic in the Q-methodology, is the size of the P-sample (see subsection 3.2). It is important to note that this is not a limitation of the Q-methodology per se as this research approach works well with small samples (see subsection 3.2). However, a small sample constitutes a limitation regarding the generalizability of the findings (see subsection 3.1). Specifically, Q-methodology studies posit that the findings consist of configurations of ideas demonstrably existent in the broader population (Brown 1993). However, further research should seek to account for the diffusion of the configurations of ideas found in this study. Thus, further research can develop a survey based on the results of this thesis and administer it to a broader population. Research on the post-adoption phases of AI in the public sector is scant because AI projects that fit under that category are scarce (see subsection 2.2). Further research may be conducted after progress is made in the implementation of the projects addressed in this thesis. The findings of such research can be compared with the findings of this thesis to provide an account of whether Estonian civil servants' configurations of ideas on AI have changed with the advancements in its use and, if so, how. Importantly, exploring the change of ideas on a given topic by conducting research based on Q-methodology is one of the uses foresought by its developers (Brown 1993). This strand of research appears to be particularly fruitful. Indeed, further research may be developed after projects aimed at solving the problems related to the possible mismatch between trust and trustworthiness highlighted by Zhu et al. (2021). In this sense, further research on trust in AI among Estonian civil servants can be used to gather insights into the effectiveness of such projects by grasping participants' perceptions on the topic. Namely, using qualitative research methods suitable for longitudinal studies in combination with Q-methodology adds longitudinality whereas the Q-methodology only provides a snapshot (Watts and Stenner 2005). In this regard, further research can be conducted in the same setting by applying qualitative methods suitable for longitudinal studies to explore the formation of configurations found in this thesis. For instance, the findings of this thesis appear consistent with Criado et al.'s (2020) conclusions on the usefulness of co-development and framing of AI as a supporting tool for the acceptance and use of AI systems by first-level bureaucrats. Similarly, the findings of this thesis seem to confirm Van Noordt and Misuraca's (2020) conclusions on framing the use of AI as a solution for environmental triggers for the effective acceptance and implementation of AI systems. From this perspective, it could be fruitful to explore the AI projects in healthcare and social affairs with a focus on such aspects and explore the formation of such ideas in these contexts.

6 Conclusions

This research draws on extant knowledge on the use of Artificial Intelligence in the public sector and aims to fill a gap in this realm by examining the subjectivity of Estonian civil servants about their trust in Artificial Intelligence systems used in the public sector. Specifically, this research aims to answer the following research questions:

R. Q. 1 *What are the configurations of ideas about trust in Artificial Intelligence systems used in the public sector among Estonian civil servants?*

R. Q. 2 *How do Estonian civil servants configure ideas about trust in Artificial Intelligence with the pre-adoption phases of Artificial Intelligence in the public sector?*

R. Q. 3 *How do Estonian civil servants configure ideas about trust in Artificial Intelligence with the post-adoption phases of Artificial Intelligence in the public sector?*

In order to answer the research question, this thesis adopts the Q-Methodology. The methodology was born in psychology and is best suited to study the subjectivity of individuals and elicit ideal-types of ideas on a given topic as they appear in the population under study (Brown 1993). Thus, Q-Methodology allows for examining the subjectivity of Estonian civil servants on the use of Artificial Intelligence in the public sector by eliciting the configuration of ideas of Estonian civil servants on this subject. Such configurations constitute ideal-types interconnection of ideas of Estonian civil servants about trust in AI within its use in the Estonian public sector. Factor analysis based on the centroid method is used to emphasize individuals' subjectivity compared to normal factor analysis (Watts and Stenner 2005). Factor analysis based on the centroid method extracts three factors. Factor rotation based on judgmental rotation is performed on Factor 1 and Factor 3 according to the threshold of eigenvalue > 1. However, after factor rotation also Factor 2 is included in the interpretation of factors configurations. Factor 2 eigenvalue is equal to 0,99. Nonetheless, the choice of keeping Factor 2 in interpreting factors' viewpoints is justified by the insightful perspective grasped by this factor. This is consistent with the prerogatives of Q-methodology that prioritize insightfulness over statistical validity and relevance (Watts and Stenner 2005). However, it is important to notice that Factor 2 accounts for only two participants, thereby making the results on this factor less relevant than the ones on Factors 1 and 3. Factors 2 and 3 are bipolar; hence, they hold opposite views. Due to this characteristic, Factor 2 and Factor 3 are split into Factor 2a and 2b, and Factor 3a and 3b. This research answers the research question through the configurations of ideas, hence the factors resulting from factor analysis. Said configurations are named: the Aware Pragmatist, the Human-Centred Innovator, the Rationalist Executor, the Heedful Explorer, and the Critical Examinator. The Aware

Pragmatist moves from a general favorable disposition toward technology and values a combination of safety mechanisms and functionality in its trust in AI. The Human-Centred Innovator builds on the readiness of smart technology to address societal issues and makes the role of leaders and professionals in ensuring that AI is harnessed for the same goal central to its trust in AI. The Rationalist Executor draws from a demystified perception of technology and centers its trust in AI on a combination of functionality and understanding of the internal processes. The Heedful Explorer moves from an equilibrated consideration of benefits and threats of technology and values safety mechanisms, overseeing professionals, and reliability and functionality of the technology in its trust in AI. The Critical Examinator disregards the technology in general and safety mechanisms and centers its trust in AI on understanding internal processes. All configurations aside from the Critical Examinator express a generally favorable disposition towards technology. The findings on the Aware Pragmatist, the Human-Centred Innovator, and the Rationalist Executor are consistent with the literature on the matter. Conversely, the findings on the Heedful Explorer and the Critical Examinator are inconsistent. Literature highlights the role of environmental and individual context on this dimension of trust in technology (Lee and See 2004; McKnight et al. 2011). All participants belong to the same environmental context, namely the Estonian public sector. However, a pattern on individual context among participants loading on the Heedful Explorer and the Critical Examinator cannot be spotted that suggests an explanation for the unfavorable disposition towards technology expressed by this factor viewpoint. Configurations of ideas on the institutional-based dimension of trust in technology seem consistent with the literature on the matter. Literature highlights the role of previous experience with a generic type of technology that the specific type is an instance of and the existence of a system of support, safety, and guarantee mechanisms in this dimension (Lee and See 2004; McKnight et al. 2011). Additionally, drawing on Zhu et al. (2021), people in the organization are included as a referent in this dimension. The configurations are consistent with the literature on this dimension in that participants' previous experiences, beliefs, and preferences are consistent with whether elements of the institutional-based trust in technology are valued or not. Importantly, this relation holds regardless of the positive or negative experience of participants loading on a given configuration with such elements. On the dimension of trust in the specific technology - hence, Artificial Intelligence systems for decision-support- the findings of this research seem consistent with the literature on the matter. Trusting beliefs in a specific technology as conceptualized by McKnight et al. (2021), are compared and combined with the conceptualization of trust in automated systems proposed by Lee and See (2004) -i.e., performance, process, and purpose. The configuration of ideas appears consistent with the literature in that a relationship between the experiences, preferences, and elements of

trust in the specific technology valued by such viewpoints can be observed (Lee and See 2004; McKnight et al. 2011; Zhu et al. 2021). As for the dimension of institutional-based trust, this holds regardless of whether participants loading on a configuration commented positively or negatively on their experience, preference, or beliefs on the elements of trust in the specific technology. Additionally, the results of this thesis shed light on the configurations of ideas of Estonian civil servants about trust in AI concerning certain influencing factors in the pre- and post-adoption phases of AI in the public sector. This constitutes meaningful insights on the subjectivity of relevant stakeholders of the use of AI in the public sector on facilitating and constraining factors. Nonetheless, these results are collateral to the findings of this thesis and constitute insights. Hence they shall be taken as such. With regards to this, insights gathered on this matter suggest that civil servants' perspectives are largely consistent with existing literature (Dwivedi et al. 2021; Wirtz et al. 2019). This is a relevant insight because literature that includes the perspective of civil servants on this matter is scarce (Ahn and Chen 2021). Moreover, the literature that includes such stakeholders mostly focuses on mid or senior-level managers (Sun and Medaglia 2019), whereas this research also includes the perspective of civil servants on lower levels. In summary, results from this thesis suggest that the views of Estonian civil servants are consistent with the literature with regard to environmental factors for the adoption of AI, such as framing of AI as a solution to external triggers (Van Noordt and Misuraca 2020), and the perception of a positive impact of AI on public value creation as influencing the implementation (Van Noordt and Misuraca 2020); the role of policies and data sharing agreements or regulations as an influencing factor for both adoption and implementation (Campion et al. 2020); management support and organizational resources as well as organizational processes in the adoption phases (Van Noordt and Misuraca 2020). This thesis does not come without limitations. On the methodological side, the concourse was developed largely based on scientific research, resulting in a Q-sample made of statements that were somewhat difficult to understand for the participants. Additionally, some statements might consist of facts, which goes against the prerogatives of Q-methodology (Brown 1993). Q-methodology is best conducted by developing a purposeful sample (Watts and Stenner 2005). However, the P-sample of this research is quite unbalanced, particularly so in the gender and field of the public services. With regards to the findings the main shortcoming is caused by the low number of participants loading on Configuration 3 and more so on Configuration 2. However, Configuration 3 is relevant from a statistical perspective whereas this is not true for Configuration 2. Additionally, the research methodology adopted in this thesis allows for an account of existing viewpoints among the population under study (Watts and Stenner 2005). However, it does not allow generalizability, nor is it suitable for conclusions on the evolution of such viewpoints over time (Watts and Stenner 2005). The former constitutes

a major limitation of this research in that it hampers the validity of the results for such configurations. The latter impacts this research in that the findings can only account for current viewpoints. Future research should focus on the latter limitations as there is high potential in combining the results of this thesis with additional research. Scholars can develop surveys based on this thesis to seek for generalization. Alternatively, additional studies retracing the approach of this thesis can be conducted to highlight differences and similarities in the views of Estonian civil servants on Artificial Intelligence in the future. Additionally, combining this approach with more suited qualitative approaches can add longitudinality to the knowledge on the viewpoints of Estonian civil servants on Artificial Intelligence.

References

Ahn, M. J., and Chen, Y.-C. 2021. 'Digital transformation toward AI-augmented public administration: The perception of government employees and the willingness to use AI in government', *Government Information Quarterly*, p. 101664.

Alshahrani, A., Dennehy, D., and Mäntymäki, M. 2021. 'An attention-based view of AI assimilation in public sector organizations: The case of Saudi Arabia', *Government Information Quarterly*, p. 101617.

Aoki, N. 2020. 'An experimental study of public trust in AI chatbots in the public sector', *Government Information Quarterly* (37:4), p. 101490.

Bailey, D. E., and Barley, S. R. 2020. 'Beyond design and use: How scholars should study intelligent technologies', *Information and Organization* (30:2), p. 100286.

Ballester, O. 2021. 'An Artificial Intelligence Definition and Classification Framework for Public Sector Applications', in *DG.O2021: The 22nd Annual International Conference on Digital Government Research*, Omaha NE USA: ACM, pp. 67–75.

Banasick, S. 2019. 'KADE: A desktop application for Q methodology', *Journal of Open Source Software* (4:36), p. 1360.

Bovens, M., and Zouridis, S. 2002. 'From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control', *Public Administration Review* (62:2), pp. 174–184.

Brown, S. R., Baltrinic, E., and Jencius, M. 2018. 'From Concourse to Q Sample to Testing Theory •', *Operant Subjectivity: The International Journal of Q Methodology*, pp. 1–17.

Brown, S. R. 1993. 'A Primer on Q Methodology', *Operant Subjectivity: The International Journal of Q Methodology* (16:3/4), pp. 91–138.

Bullock, J., Young, M. M., and Wang, Y.-F. 2020. 'Artificial intelligence, bureaucratic form, and discretion in public service', *Information Polity* (25:4), S. Giest and S. Grimmelikhuijsen (eds.), pp. 491–506.

Campion, A., Gasco-Hernandez, M., Jankin Mikhaylov, S., and Esteve, M. 2020. 'Overcoming the Challenges of Collaboratively Adopting Artificial Intelligence in the Public Sector', *Social Science Computer Review*, pp. 1–16.

Campion, A., Hernandez, M.-G., Mikhaylov Jankin, S., and Esteve, M. 2020. 'Managing Artificial Intelligence Deployment in the Public Sector', *Computer* (53:10), pp. 28–37.

Criado, J. I., Valero, J., and Villodre, J. 2020. 'Algorithmic transparency and bureaucratic discretion: The case of SALER early warning system', *Information Polity* (25:4), S. Giest and S. Grimmelikhuijsen (eds.), pp. 449–470.

Deley, T., and Dubois, E. 2020. 'Assessing Trust Versus Reliance for Technology Platforms by Systematic Literature Review', *Social Media + Society* (6:2), pp. 1–8.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., et al. 2021. 'Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy', *International Journal of Information Management* (57), p. 101994.

Dziopa, F., and Ahern, K. 2011. 'A Systematic Literature Review of the Applications of Q-Technique and Its Methodology', *Methodology* (7:2), pp. 39–55.

Gregor 2006. ‘The Nature of Theory in Information Systems’, *MIS Quarterly* (30:3), pp. 611–642.

Guenduez, A. A., Mettler, T., and Schedler, K. 2020. ‘Technological frames in public administration: What do public managers think of big data?’, *Government Information Quarterly* (37:1), pp. 1–12.

Heeks, R., and Bailur, S. 2007. ‘Analyzing e-government research: Perspectives, philosophies, theories, methods, and practice’, *Government Information Quarterly* (24:2), pp. 243–265.

Horowitz, M. C., and Kahn, L. 2021. ‘What influences attitudes about artificial intelligence adoption: Evidence from U.S. local officials’, *PLOS ONE* (16:10), C. Delcea (ed.), pp. 2–20.

Kumar, S., Raut, R. D., Queiroz, M. M., and Narkhede, B. E. 2021. ‘Mapping the barriers of AI implementations in the public distribution system: The Indian experience’, *Technology in Society* (67), p. 101737.

Kuziemski, M., and Misuraca, G. 2020. ‘AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings’, *Telecommunications Policy* (44:6), p. 101976.

Lee, J. D., and See, K. A. 2004. ‘Trust in Automation: Designing for Appropriate Reliance’, *Human Factors* (46:1), pp. 50–80.

Lorenz, L., Meijer, A., and Schuppan, T. 2021. ‘The algocracy as a new ideal type for government organizations: Predictive policing in Berlin as an empirical case’, *Information Polity* (26), pp. 71–86.

McKnight, D. H., Carter, M., Thatcher, J. B., and Clay, P. F. 2011. ‘Trust in a specific technology: An investigation of its components and measures’, *ACM Transactions on Management Information Systems* (2:2), pp. 1–25.

Meijer, A., Lorenz, L., and Wessels, M. 2021. ‘Algorithmization of Bureaucratic Organizations: Using a Practice Lens to Study How Context Shapes Predictive Policing Systems’, *Public Administration Review* (81:5), pp. 837–846.

Mikalef, P., Lemmer, K., Schaefer, C., Ylinen, M., Fjørtoft, S. O., Torvatn, H. Y., et al. 2021. ‘Enabling AI capabilities in government agencies: A study of determinants for European municipalities’, *Government Information Quarterly*, p. 101596.

Miller, T. 2019. ‘Explanation in artificial intelligence: Insights from the social sciences’, *Artificial Intelligence* (267), pp. 1–38.

Naderifar, M., Goli, H., and Ghaljaie, F. 2017. ‘Snowball Sampling: A Purposeful Method of Sampling in Qualitative Research’, *Strides in Development of Medical Education* (14:3).

Newman, J., Mintrom, M., and O’Neill, D. 2022. ‘Digital technologies, artificial intelligence, and bureaucratic transformation’, *Futures* (136), p. 102886.

Paige, J. B., and Morin, K. H. 2014. ‘Q-Sample Construction’, *Western Journal of Nursing Research* (38:1), pp. 96–110.

Peeters, R. 2020. ‘The agency of algorithms: Understanding human-algorithm interaction in administrative decision-making’, *Information Polity* (25), pp. 507–522.

Pencheva, I., Esteve, M., and Mikhaylov, S. J. 2020. ‘Big Data and AI – A transformational shift for government: So, what next for research?’, *Public Policy and Administration* (35:1), pp. 24–44.

Pi, Y. 2021. ‘Machine learning in Governments: Benefits, Challenges and Future Directions’, *JeDEM - eJournal of eDemocracy and Open Government* (13:1), pp. 203–219.

Robbins, B. G. 2016. ‘What is Trust? A Multidisciplinary Review, Critique, and Synthesis’, *Sociology Compass* (10:10), pp. 972–986.

Samoili, S., Lopez Cobo, M., Gomez Gutierrez, E., De Prato, G., Martinez-Plumed, F., and Delipetrev, B. 2020. ‘AI watch: defining Artificial Intelligence : towards an operational definition and taxonomy of artificial intelligence.’, *Publications Office of the European Union*.

Sousa, W. G. de, Melo, E. R. P. de, Bermejo, P. H. D. S., Farias, R. A. S., and Gomes, A. O. 2019. ‘How and where is artificial intelligence in the public sector going? A literature review and research agenda’, *Government Information Quarterly* (36:4), p. 101392.

Sun, T. Q., and Medaglia, R. 2019. ‘Mapping the challenges of Artificial Intelligence in the public sector: Evidence from public healthcare’, *Government Information Quarterly* (36:2), pp. 368–383.

Thatcher, J. B., McKnight, D. H., Baker, E. W., Arsal, R. E., and Roberts, N. H. 2011. ‘The Role of Trust in Postadoption IT Exploration: An Empirical Examination of Knowledge Management Systems’, *IEEE Transactions on Engineering Management* (58:1), pp. 56–70.

Thiebes, S., Lins, S., and Sunyaev, A. 2021. ‘Trustworthy artificial intelligence’, *Electronic Markets* (31:2), pp. 447–464.

Van Noordt, C., and Misuraca, G. 2022. ‘Artificial intelligence for the public sector: results of landscaping the use of AI in government across the European Union’, *Government Information Quarterly* (39:3), p. 101714.

Van Noordt, C., and Misuraca, G. 2020. ‘Evaluating the impact of artificial intelligence technologies in public services: towards an assessment framework’, in *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, Athens Greece: ACM, pp. 8–16.

Van Noordt, C., and Misuraca, G. 2020. ‘Exploratory Insights on Artificial Intelligence for Government in Europe’, *Social Science Computer Review*, pp. 1–19.

Van Noordt, C. 2022. ‘Conceptual challenges of researching Artificial Intelligence in public administrations’, Seoul Republic of Korea: ACM, pp. 1–8.

Veale, M., and Brass, I. 2019. ‘Administration by Algorithm?’, in *Algorithmic Regulation*, Oxford: Oxford University Press, pp. 1–30.

Vydra, S., and Klievink, B. 2019. ‘Techno-optimism and policy-pessimism in the public sector big data debate’, *Government Information Quarterly* (36:4), p. 101383.

Wang, Y., Zhang, N., and Zhao, X. 2020. ‘Understanding the Determinants in the Different Government AI Adoption Stages: Evidence of Local Government Chatbots in China’, *Social Science Computer Review*, p. 089443932098013.

Watts, S., and Stenner, P. 2005. ‘Doing Q ethodology: theory, method and interpretation’, *Qualitative Research in Psychology* (2:1), pp. 67–91.

Wirtz, B. W., Langer, P. F., and Fenner, C. 2021. ‘Artificial Intelligence in the Public Sector - a Research Agenda’, *International Journal of Public Administration* (44:13), pp. 1103–1128.

Wirtz, B. W., Weyerer, J. C., and Geyer, C. 2019. ‘Artificial Intelligence and the Public Sector—Applications and Challenges’, *International Journal of Public Administration* (42:7), pp. 596–615.

Young, M. M., Bullock, J. B., and Lecy, J. D. 2019. ‘Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration’, *Perspectives on Public Management and Governance* (2:4), pp. 301–313.

Zachariadis, M., Barrett, M., and Scott, S. 2013. ‘Methodological Implications of Critical Realism for Mixed-Methods Research’, *MIS Quarterly* (37:3), pp. 855–879.

Zhang, W., Zuo, N., He, W., Li, S., and Yu, L. 2021. ‘Factors influencing the use of artificial intelligence in government: Evidence from China’, *Technology in Society* (66), p. 101675.

Zhu, L., Xu, X., Lu, Q., Governatori, G., and Whittle, J. 2021. ‘AI and Ethics -- Operationalising Responsible AI’, *arXiv:2105.08867 [cs]*.

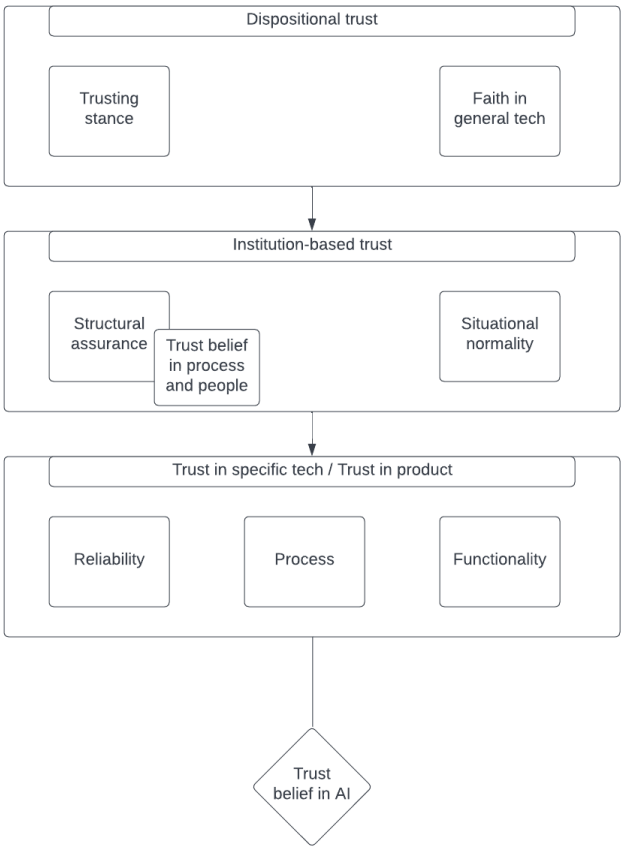
Zuiderwijk, A., Chen, Y.-C., and Salem, F. 2021. ‘Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda’, *Government Information Quarterly* (38:3), p. 101577.

Appendix

A Resources for the Concourse

Resource	Author	Sector
Research Paper	Ahn and Chen	Academia
Research Paper	Alsharani et al.	Academia
Research Paper	Campion et al.	Academia
Research Paper	Criado and Zarate-Alcarazo	Academia
Research Paper	Goad and Gal	Academia
Research Paper	Mikalef et al.	Academia
Research Paper	Peeters	Academia
Research Paper	Sun and Medaglia	Academia
MSc Thesis	Vrieling	Academia
Publication	Deloitte	Private Sector
Conference	InTouchAI.eu	Third Sector
Conference	AI4Belgium	Third Sector
Publication	World Economic Forum	Third Sector
Lecture	European Trade Union Institute	Third Sector

B Conceptual Map



C Q-Sample

STATEMENT NUMBER	STATEMENT
1	Technology is still not always ready to address some concerns of societal benefits
2	The fundamental problem is that technology simplifies the world transforming it into numbers and then feeding them back into reality
3	Development of technology and computation of data for and by AI are a cause of environmental and social issues
4	ADM can be trusted
5	AI makes it possible to prove discrimination and bias in decision-making, whereas this is barely possible for human decision-making
6	The large-scale deployment typical of AI systems is challenging as it prompts a number of tradeoffs
7	The economic value of AI outweighs the concerns
8	AI is a competitor to humans at work
9	New professions in the public sector linked to AI will have to be tackled thoroughly and other “intelligences” or capabilities will be required from civil servants.
10	Informal relationships are important in the process of setting up a data-sharing agreement. I think that showing that you have ideas, and not waiting until that agreement is signed, to get things kick-started, can be helpful too. More often than not, you have to build some trust before you start with the institutional agreements. Then once you have that formal agreement in place, that’s a symbol.
11	There are people who have no clue where to begin. Either because they have no data literacy, or because no one has documented what this is supposed to mean; or because someone else decided this is the indicator of success, and they don’t necessarily believe that. Moreover, developing AI projects in some cases requires that people in organizations have a basic understanding of what a data-oriented question is and how it can be solved using an AI technique.
12	Reliability of AI system comes from the engagement of people affected by the system
13	It is important to know that the information officers that are working with the system are qualified people
14	Ethical AI needs to be ingrained in organizational culture just like the mission and vision of agencies.
15	Agencies need a leader responsible for keeping a focus squarely on ethical AI practices and fostering the coordination to make it happen
16	Responsibility is the key and it must be addressed individually through ethical code of conducts for developers
17	Reporting and advice channels are important to build a transparent, trustworthy and inclusive environment where gathering info on AI ethics is normal practice, for instance there can be AI ethics hotlines
18	AI systems need policies about who is responsible and accountable for their output or decision-making.
19	what I don’t want to do is get tied up in information-sharing agreements and end up with issues you can’t resolve. For instance, sensitive information that the police might hold, they might not want to share with a local authority and vice versa
20	Cross-sectoral cooperation is essential for trustworhty AI systems
21	Purpose-oriented auditing for ethics is needed
22	Focusing on the human in the loop at the individual level is not enough to ensure beneficial AI for humans because it doesn't account for the network effect
23	Technical standards are the main source to promote and safeguard ethical values
24	AI is an asset to deliver on the organization's mission
25	Selflearning can lead to improved accuracy and quality of decision-making
26	It is difficult to follow-up on AI predictions because they can generate accurate but counterintuitive insights due to the large number of variables and data they use. They go against the grain of traditional heuristics. They challenge the ways things have traditionally been done. And they often require people to give up familiar tools and methods.
27	As the quality of automation improves it can be assumed that the system is infallible

28	AI systems making decisions can be deployed at scale, possibly affecting millions of stakeholders. Any inaccuracies can translate into a large number of mistakes.
29	Accuracy is more important than an explainable system
30	Not only AI systems may change over time but also the context, this can make the performance deteriorate
31	Stability is particularly important in the public sector, where many external factors affect decision-making.
32	Machines can process more information at a higher speed than even the best trained humans are capable of. There is no way in which the human operator can check in real-time that the computer is following its rules correctly.
33	Explanations on accuracy are important
34	The fact that AI can learn negative behavior and be manipulated by humans is worrysome
35	Algorithms may be used as a convenient default for human decision-making, thereby reflecting wellknown mechanisms of satisficing behaviour
36	Automation is “most dangerous when it behaves in a consistent and reliable manner most of the time”. In accordance, especially in the case of routinized tasks, human decision-makers at screen-level are likely to face difficulties in identifying abnormalities that warrant the use of their discretion.
37	AI systems make decisions often using complex mathematical equations. Despite impressive scientific progress, it is not possible to guarantee that the decisions these systems make are always ethical, especially when they operate on new data and possibly in changing environments once they are deployed. For example, there is no guarantee that an AI system will be fair across different groups of people over time, even if it has been tested and shown to be so using previous data.
38	Moreover, because these systems can consist of multiple components possibly developed by various organizations and based on data across different sources, it can be impossible to identify responsibilities for errors. This can hinder accountability.
39	Agencies should emphasize creation of algorithms that are transparent and can be explained to people who are being impacted by those algorithms.

D Fixed Quasi-Normal Distribution