Quaderni di Comunità

Persone, Educazione e Welfare
nella società 5.0

Community Notebook

People, Education, and Welfare in society 5.0

n. 1/2025

HUMAN-CENTRIC APPROACH
TO ARTIFICIAL INTELLIGENCE

edited by
Marco Filoni, Filippo Maria Giordano, Giorgio Grimaldi

eurilink
University Press

Iscrizione presso il Registro Stampa del Tribunale di Roma
al n. 172/2021 del 20 ottobre 2021

Eurilink University Press Srl
Via Gregorio VII, 601 - 00165 Roma
www.eurilink.it - ufficiostampa@eurilink.it
ISBN: 979 12 80164 90 2
ISSN: 2785-7697 (Print)
ISSN: 3035-2525 (Online)

Prima edizione, luglio 2025
Progetto grafico di Eurilink

INDICE

5. A MODEL FOR RESPONSIBLE GOVERNANCE OF HUMAN-CENTRIC AI IN THE PUBLIC SECTOR

by Francesco Niglia*

Abstract: La governance responsabile dell'Intelligenza Artificiale nel settore pubblico non è più un'opzione rimandabile a causa dei numerosi problemi etici emersi negli ultimi anni caratterizzati da una massiccia adozione di servizi basati sull'IA. Date le numerose sfide poste, diviene obbligatorio includere prospettive sociali incentrate sull'uomo. Questo studio discute un modello di framework per definire i ruoli, le responsabilità e le competenze di tutti gli stakeholder coinvolti nei processi di sviluppo, distribuzione e valutazione dell'IA nel settore pubblico.

Parole chiave: governance, responsabilità, affidabilità, modello, pubblica amministrazione, antropocentrismo.

Abstract: Responsible AI Governance in the Public Sector is no longer an option due to the numerous ethical issues that have emerged in recent years with the adoption of AI-based services in the Public Sector. Given the numerous challenges AI poses, it is essential to incorporate human-centric and social perspectives. This study discusses a framework model for defining the roles, responsibilities, and skills of all the stakeholders involved in the processes of AI development, deployment, and assessment.

* Università degli Studi "Link", https://orcid.org/0000-0001-6452-5189, f.niglia@unilink.it.

Accettato Dicembre 2024 - Pubblicato Aprile 2025.

Keywords: governance, responsibility, trustworthiness, framework, public sector, human-centric.

Introduction

This essay builds upon theoretical research to develop a framework for implementing responsible governance of Artificial Intelligence in the public sector and identifying the minimum skills necessary to achieve this goal. The core research question lies in providing evidence about the operational feasibility of EU trustworthy AI guidelines by deeply analysing ethical and responsibility requirements and connecting the dots between the theoretical approaches and early practice of AI governance examples. The study draws on official European Commission reports on the ethical deployment and use of AI and several existing frameworks proposed by renowned research teams and international innovation agencies that aim to implement social and ethical AI. The essay's strategy examines how responsibilities among stakeholders are being reconfigured to achieve new equilibria necessitated by introducing AI-based solutions in the public sector and consequently outlines a possible framework of roles and competencies.

1. What is Responsible Human-Centric AI Governance, and why is it relevant

Artificial Intelligence Governance in the Public Sector is not more of an option since AI has become a «key strategic asset of public service in the last few years» (Tangi L. *et al.*, 2022), and the whole Public Sector witnessed it through its rapid engagement in

the government's AI appropriation procedures (European Commission (2024a)). The capacity of AI technologies and systems to operate in ways that were not previously possible may challenge our existing legal, moral, and social conceptions of responsibility, particularly given their inner properties. The deployment of AI systems has given many examples of how it could impact or infringe on human rights and human-centred values, decreasing or excluding human needs, values, and capabilities. Such characteristics leverage the capacity to generate 'hidden' insight from merging data sets, the ability to accurately imitate human traits, greater software complexity, inscrutability and opacity, and the risk of generating collective action problems. AI ethics is, therefore, required as an integral component due to its potential to touch many aspects of society. Policymakers and all actors in AI should implement safeguards to address significant risks arising from purposeless, intentional, or unintentional misuses. Although AI development has quickly outpaced the speed of regulation, policymakers and regulators have started fining companies for either developing or simply using biased AI algorithms (Hak, A., 2022). As a further confirmation that Ethics and Responsibility in AI development are a global issue, we can list the growing number of international standards highlighting a broad and shared consensus for ensuring the adoption of responsible AI. Starting from the EU AI Act in 2019, examples are the OECD AI Principles and the UNESCO Recommendations on the Ethics of AI in 2021, the G7 Toolkit for AI in the Public Sector, and the UN Resolution in 2024 (UN, 2024). Responsible Human-centric AI Governance (selected acronym RHAIG for this study's purposes) constitutes a common requirement in this international scenario; however, despite being a simple concept, it reveals a pretty complex structure behind it because it harmonises Responsibility, Trustworthiness, and Human-Centric approaches. The first step in

untangling the “conceptual knots” is to clarify the differences among these terms. Responsibility is the process of defining policies and establishing accountability to guide the creation and deployment of AI systems in an organisation. Its scope and criteria have been determined by the High-Level Expert Group on Artificial Intelligence (AI-HLEG, 2019). It leverages four ethical principles and seven requirements (details in Table 1): ethical principles are imperatives rooted in fundamental rights, which must be respected to ensure that AI systems are developed, deployed and used in a trustworthy and human-centric manner; Key Requirements can translate the principles into concrete actions to achieve Trustworthy AI; they include systemic, individual and societal aspects and apply to different stakeholders in AI systems’ life cycle: developers, deployers, end-users, and the broader society. Trustworthiness in AI ensures fairness, inclusion, and benefit for all members of society. It is central to the future acceptance and adoption of AI technologies in all public areas: the realisation of trusted technologies requires cross-cutting collaboration in governance (law, regulation, ethics), technology (explainability, uncertainty, fairness/bias), and end-user application domains (health, business, entertainment, etc.). Human-centred AI prioritises human needs, values, and capabilities, as suggested by the OECD Human-Centred Values and Fairness Principle: «AI should be developed in a way that is consistent with human-centred values, such as fundamental freedoms, equality, fairness, rule of law, social justice, data protection and privacy, and consumer rights and commercial fairness» (OECD, 2021). The primary assumption of this study is to consider trustworthiness and human-centric principles as grounding elements of Responsibility in AI governance, assuming that responsible AI governance is the framework that allows respecting trustworthy AI principles.

Table 1 presents the requirements and related criteria in detail, linking the principles to the operational domain. In this

scheme, addressing information for all AI stakeholders designing, developing, deploying, implementing, using, or being affected by AI is compelling. The list includes (but is not limited to) companies, organisations, researchers, public services, government agencies, institutions, civil society organisations, individuals, workers and consumers.

Table 1: Details of requirements and criteria for Responsible Human-Centric AI Governance

Principle	*Key Requirement*	*Criteria*	*Main AI-related threat*
Respect for human autonomy	Human agency and oversight	Fundamental rights	Given the reach and capacity of AI systems, they can also negatively affect fundamental rights.
		Human agency	Users should be able to make informed autonomous decisions regarding AI systems.
		Human oversight	It helps ensure that an AI system does not undermine human autonomy or cause adverse effects.
Prevention of harm	Technical robustness and safety	Resilience to attack and security	AI systems […] should be protected against vulnerabilities that can allow them to be exploited by adversaries.
		Fallback plan and general safety	AI systems should have safeguards that enable a fallback plan in case of problems.
		Accuracy	It pertains to an AI system's ability to make correct judgments.
		Reliability and Reproducibility	It is critical that the results of AI systems are reproducible and reliable.
	Privacy and data governance	Privacy and data protection	AI systems must guarantee privacy and

			data protection throughout a system's entire lifecycle (GDPR).
		Quality and integrity of data	The quality of the data sets used is paramount to the performance of AI systems.
		Access to data	Only duly qualified personnel with the competence and need to access an individual's data should be allowed to do so
	Societal and environmental well-being	Sustainable and environmentally friendly AI	AI systems promise […] must be ensured that this occurs in the most environmentally friendly way possible.
		Social impact	AI systems can contribute to the deterioration of social skills and affect people's physical and mental well-being.
Fairness		Society and Democracy	The use of AI systems should be carefully considered in situations involving the democratic process, including political decision-making and electoral contexts.
Explicability	Transparency	Traceability	It facilitates auditability as well as explainability
		Explainability	Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings.
		Communica-tion	Humans have the right to be informed that they are interacting with an AI system.
Fairness	Diversity, non-discrimination, and fairness	Avoidance of unfair bias	Data sets used by AI systems (both for training and operation) may suffer from the inclusion of

			inadvertent historical bias, incompleteness, and bad governance models.
		Accessibility and universal design	Systems should be user-centric and designed to allow all people to use AI products or services, regardless of their age, gender, abilities, or characteristics.
		Stakeholder Participation	It is advisable to consult stakeholders who may directly or indirectly be affected by the system throughout its life cycle.
	Accountability	Auditability	It entails enabling the assessment of algorithms, data, and design processes.
		Minimisation and reporting of negative impacts	The ability to report on actions or decisions that contribute to a certain system outcome and respond to the consequences of such an outcome must be ensured.
		Trade-offs	Tensions may arise between requirements, leading to inevitable trade-offs that should be addressed rationally and methodologically within the state of the art.
		Redress	Accessible mechanisms should be foreseen to ensure adequate redress.

Source: author's elaboration of the main AI-related threats based on information and classification provided in the "Ethics Guidelines for Trustworthy Artificial Intelligence by the High-Level Expert Group on AI", 2019.

2. Methodology of this study

After having broadened the concepts of responsible human-centric AI and trustworthiness, the study follows the logical steps:

- Build a detailed list of features and boundaries of responsible governance for the public sector. Building upon the requirements for RHAIG, the study evaluates current solutions and examples of public services and responsible governance made with and by AI. These paradigms contribute to building a knowledge base about what the model should answer AI's expected and potential role.
- Define a general-purpose qualitative model for RHAIG. The model embeds all the grounding elements of responsible and human-centric AI governance in the public sector. To this end, the author evaluates frameworks, methods, and concepts that can take responsibility and operational tasks by performing grey and academic literature analysis on existing AI governance frameworks. The academic survey considers the existing scientific articles published in Scopus and found through the query "*responsible* + *human* + *artificial intelligence* + *governance* + *framework*."
- The grey analysis asks for public information from private organisations and agencies; the author used the same query to find reports, white papers, recommendations, and policy statements. Once selected for their relevance, all the documents are analysed to find differences and commonalities in the approach to AI governance. The RHAIG model includes the most relevant features of existing frameworks; all the elicited knowledge results in a general model (Figure 2) and requirement mapping (Table 2).

- Finally, a matching exercise helps start assigning a minimum set of skills and competencies to ensure the implementation of responsible human-centric AI deployments in the Public Sector. The author deploys a recent EU report on AI for Public Sector skills and procedural framework to evaluate which competencies are suitable for each responsibility and role's layers in RHAIG (Table 3).

3. The Responsible Appropriation of AI in the Public Sector

Adopting AI in the public sector forces EU public administrations to «face the challenges of implementing AI solutions» (Grimmelikhuijsen, S. Tangi, L., 2024). A recent EC Public Sector Tech Watch observatory report mapped over a thousand AI cases in Europe, and its findings evidence the difference in the adoption of AI between the national and local levels. It identified a substantial difference: the former case is more interested in «improving processes and bureaucratic tasks». At the same time, the latter mainly uses AI for «enhancing the implementation of public services» (European Commission, 2024b). Given this difference, the author opted for a general approach to avoid territorial levels and consider three aspects of the process: the challenges, the governance layers, and the responsibility assignment. Trustworthy AI systems are expected to contribute to the public good; in both local and national contexts, all the potential solutions to AI governance shall consider and coherently balance the effort to «prevent possible harm that AI solutions could have to the individual and the community / collective levels» (Andrus, M., & Villeneuve, S., 2022). Individual risks relate to Privacy, Bias, Identity Misrepresentation, and Data Misuse. Collective risks consider the following: Expanded Surveillance

(invasive for targeted categories), Misrepresentation of Categories (misalignment with identity or experience), and Private Control for Fairness goal (potential blind spots in unfairness). Mindful of the opportunities of AI in the Public Sector and leveraging the capability to «enhance or optimise existing public services and create new services» (Manzoni M., *et al.*, 2022), public administrations started asking themselves whether they could be held responsible for the adverse consequences AI systems might generate; the obligations of the state concerning AI adverse effects ascribe responsibilities to those who develop and implement AI systems. It turns out that policymakers face a dilemma: the obligation to protect citizens from potential algorithmic harms coexists with the mandate to increase efficiency and enhance the quality of digital services (Molinari F. *et al.*, 2021). Several ways exist to build coherent and solid control frameworks for Trustworthy human-centric AI governance; the most common leverage are Rights and liabilities (protection), Command and control (penalties), Administrative oversight (agencies), Incentives (tax credit), Market-harnessing controls (non-economic driven), Public infrastructure (informing values), Mandatory disclosures (performance-related metrics), Public compensation (revenues compensating for harms). In this context, policymakers have the opportunity and duty to guide the development of fair AI systems that benefit society. At this scope, the European AI Act, which formally entered into force on August 1st, 2024, aims to foster trustworthy AI in Europe by ensuring that «AI systems respect fundamental rights, safety, and ethical principles while addressing the risks of AI models» (European Commission, 2021).

3.1 Key challenges for AI governance for administration

The policymakers' challenge is twofold: to govern AI, algorithms, and related automated processes, and govern with and by AI, using algorithms and computerised methods and systems to enhance and improve public services. They constitute different perspectives brought by the pervasive character of AI-enabled processes in the public sector:

- Governing AI requires a framework or a set of rules and laws, such as data regulatory regimes and practices, to address the primary driver of AI development, the conjunction of a massive quantity of data with innovative machine learning algorithms.
- Governing with and by AI represents the effective use and value AI can offer governments when redesigning internal administrative processes to enhance the quality and impact of public services. It still requires humans to be "in the loop" by using, controlling, and supervising technology that reinforces public offerings' capacity. The implementation of AI-based public services, however, requires a deeper understanding of potential risks associated with the expected benefits.

The increased maturity of citizens' interaction with digital means helps the Public Sector provide accurate, efficient services. In this context, AI-based technologies shall be considered a component of complex socio-technical-economic systems, and they generate several challenges in identifying roles and responsibilities in service delivery and in finding balanced governance. Each approach to governance comes with benefits and downsides, so effective AI governance in the appropriation process should strategically combine them to maximise impact and minimise harm or risks. AI appropriation helps appreciate the number of different

stakeholders involved in the governance of AI-based solutions for the government. «By this term [we mean] the union of adoption and implementation, also bearing in mind that individual and group users of a certain technology after it is embedded in an organisational setup make changes to both the technology and the environment, which inevitably feed back into both steps of the singled-out process» (Molinari F. *et al.*, 2021). The first challenge becomes clear: AI intensifies inequality between people with access to its services and those without access and between those who benefit from it and those who do not, causing the "digital divide". Combining different data sources and predictive tools generates new insights into citizens, leading to unequal rights among citizen groups based on their use of AI-based services. Other challenges due to the seamless improvement and development of new digital solutions are mentioned in (Council of Europe, 2019): the problems of 'many hands', 'humans in the loop', and the 'unpredictable effects' of complex dynamics. The 'many hands' problem arises when attempting to identify the responsibilities for errors and harms resulting from the development and operation of AI systems based on several critical components. The 'Human-in-the-Loop' problem arises when considering that many tasks previously performed by humans are now partially undertaken by machines without control. The "unpredictability" problem exists when trying to identify, anticipate, and prevent adverse events arising from the interactions of two or more AI-based systems or solutions (e.g., two algorithms). When considering the process of AI appropriation, these challenges extend to the number of different stakeholders and procedures, thus introducing systemic issues and additional governance challenges: Multi-level governance between national and local levels, Accountability of political actions and automated systems, *Limited skills* among public servants, and *Limited Economic capacity* of local authorities.

3.2 The layers of AI governance within the appropriation procedures

Public policymakers need to assume a front-and-centre role in ensuring beneficial AI systems and, simultaneously, «face the challenges and overcome critical factors» (Misuraca, G., Van Noordt, C. 2020). A layered model is helpful when considering how policies can govern AI, aiming to define appropriate behaviour for AI and autonomous systems. This study deepens the application of an existing four-layer model proposed by (Gasser, U., Almeida, V. A., 2017):

- Technical layer. It is the foundation of the AI governance ecosystem, including the algorithms and data from which it is built. It establishes who is responsible for norms, regulations, and legislation.
- Ethical layer. It articulates high-level ethical concerns that apply to all types of AI applications and systems, including human rights principles introduced by trustworthy AI. It establishes who is responsible for the criteria and principles.
- Social and Legal layer. It addresses the process of creating institutions and allocating responsibilities for regulating AI and autonomous systems. One starting point for specific norms aimed at regulating AI can be the principles and criteria that emerge from the ethical and technical layers, in addition to pre-existing and more general national and international legal frameworks. It established who is responsible for data governance, algorithmic accountability, standards, and the evaluation of trade-offs in the design of AI systems.
- Evaluation layer. It covers as much of the lifecycle as possible to ensure that the focus considers productive deployment and evaluation. It establishes who is responsible for ex-ante and

continuous ex-post impact assessment and cost-benefit analysis of the acquired AI-based service.

3.3 Theoretical models for allocating responsibilities

The proposed RHAIG framework includes *Responsibility* layers. The "responsibility models" are mandatory when approaching high-risk categories identified by the AI Act; nevertheless, they are instrumental in correctly identifying how to uniquely assign responsibilities and formalities when facing systems' faults or direct negligence. A report from the Council of Europe (Yeung, K., 2019) outlines four broad heuristic responsibility models addressing stakeholders involved in AI appropriation:

- Intention/culpability-based models. They focus primarily on the voluntariness of the agent's conduct. They can be interpreted as requiring the satisfaction of the 'control' condition, demonstrating that the agent was causally responsible for the legally proscribed conduct as the agent had a free and voluntary choice concerning whether to act, notwithstanding the harmful consequences of the agent's conduct.
- Risk/negligence-based models. They refer to harm as a reasonably foreseeable consequence of the computational systems' actions and decisions, and can be ascribed to the human developers of computational agents and systems.
- Strict Responsibility. Responsibility attaches to the agent without proof of fault, so legal responsibility for rights violations attaches to those who cause them, regardless of whether the responsible agent engaged in conduct that breached a legally specified standard of conduct.

- Mandatory insurance schemes. It could be established on a 'no-fault' basis by establishing an insurance fund to which all those harmed by the operation of these technologies could have recourse. It admits to simply requiring stakeholders in the AI appropriation value chain to take out mandatory liability insurance.

No one policy option is perfect; each has its drawbacks and should be viewed within the context of an entire toolbox from which policymakers can draw. The harms of AI can be varied, and so too should the policy instruments used to address them if an effective change is to be made. However, the concept of responsibilities is a common characteristic of all the possible models and governance frameworks.

4. A Qualitative Model for Responsible Human-Centric AI Governance

4.1 Map of existing frameworks

Few researchers focus on concrete frameworks that facilitate the adoption of artificial intelligence. Recent works have compared existing literature frameworks suitable for helping to adopt AI in the public sector (Pechtor, V., & Basl, J., 2022); (Zuiderwijk, A., *et al.*, 2021). The analysis discussed the weight that seven main requirements have in each proposed framework: 1-Coverage of the AI lifecycle; 2-Concrete steps and fields of action; 3-Granularity on multiple levels; 4-Agile aspects (e.g., MLOps); 5-Consider AI maturity; 6-Incorporation of other methods (e.g., Technology–Organisation–Environment Framework, Diffusion of Innovation); 7-Inclusion of ethic or human-centric elements. In

total, 11 relevant articles suggest frameworks covering different aspects of AI in public services. The variety ranges from the prevention of discrimination, assessing the impact of AI, or even multi-layer conceptual frameworks:

- A realist perspective on AI-era public management doi.org/10.1145/3325112.3325261.
- AI innovation for advancing public service: The case of China's first administrative approval bureau. doi.org/10.1145/3325112.3325243.
- AI-Enabled Innovation in the Public Sector: A Framework for Digital Governance and Resilience doi.org/10.1007/978-3-030-57599-1_9.
- Deep learning meets deep democracy: Deliberative governance and responsible innovation in artificial intelligence. doi.org/10.1017/beq.2021.42.
- Evaluating the impact of artificial intelligence technologies in public services: Towards an assessment framework doi.org/10.1145/3428502.3428504.
- Exploring the implementation of artificial intelligence in the public sector doi.org/10.37394/232010.2020.17.9.
- Garbage in, garbage out: The vicious cycle of AI-based discrimination in the public sector doi.org/10.1145/3325112.3328220.
- Governance of artificial intelligence: A risk and guideline-based integrative framework doi.org/10.1016/j.giq.2022.101685.
- IoT and AI for smart government: A research agenda doi.org/10.1016/j.giq.2019.02.003.
- OECD Recommendation of the Council on Artificial Intelligence OECD/LEGAL/0449.

- Value-Based Guiding Principles for Managing Cognitive Computing Systems in the Public Sector doi.org/10.1080/15309576.2021.1879883.

Existing academic frameworks for the public sector are relatively few and vary in focus and application. Those covering the AI adoption process are more conceptual, while frameworks that help identify concrete fields of action are still scarce. "AI maturity" is considered in only four out of 11 proposed frameworks. As a general consideration, finding commonalities among these proposals remains difficult at the current time.

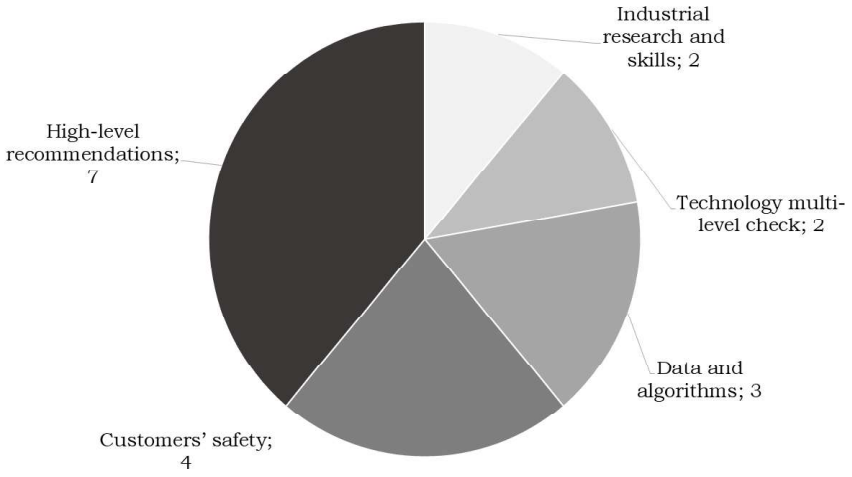
Analysing grey literature released by private and other initiatives yields usable results. The White Paper released by TNO (Veenstra, A.F. *et al.* 2021) considers 17 AI frameworks by focusing on a survey of current regulatory systems for AI and an analysis of existing ethical frameworks and methods. The author updated the list by substituting the revised OECD proposal and adding the work by Wirtz on the risks associated with public administrations. In total, the author considers 18 AI frameworks embedding ethical elements for the analysis:

- AI Guidelines (Deutsche Telekom)
- AI in the UK: ready, willing and able? – UK House of Lords, Select Committee on AI
- AI Policy Principles – Information Technology Industry Council (ITI)
- An Ethical Framework for a Good AI Society (Led by Prof. Luciano Floridi)
- Artificial Intelligence and Machine Learning: Policy Paper – Internet Society
- Automated and Connected Driving: Report – Federal Ministry of Transport and Digital Infrastructure, Ethics Commission

- Charlevoix Common Vision for the Future of Artificial Intelligence – Leaders of the G7
- Dynamic impact assessment methodology for responsible algorithmic decision-making (TNO)
- Ethically aligned design – Institute of Electrical and Electronics Engineering
- Everyday Ethics for Artificial Intelligence (IBM)
- How can humans keep the upper hand? Report on ethical matters raised by AI algorithms – French Data Protection Authority (CNIL)
- Montréal Declaration for Responsible Development of AI – Université de Montréal
- Preparing for the future of Artificial Intelligence – US National Science and Technology Council, Committee on Technology
- Principles for Accountable Algorithms and a Social Impact Statement for Algorithms – Fairness, Accountability and Transparency in Machine Learning
- The dark sides of Artificial Intelligence: An integrated AI governance framework for public administration, Wirtz (2020)
- Tools for trustworthy AI: A framework to compare implement-tation tools for trustworthy AI systems, OECD (2021).
- Top 10 Principles for Ethical Artificial Intelligence – UNI Global Union
- Z-Inspection initiative, led by Professor Roberto Zicari (Arcada University of Applied Sciences, Helsinki, Finland) and supported by several researchers worldwide.

Figure 1 summarises a brief author's at-purpose classify-cation of the 18 frameworks. Most of them are recommendations (7), followed by narrower topics such as privacy and safety (4), data use (3), technology assessment (2), and industrial research or development (2).

Figure 1: Classification of existing responsible AI frameworks by considering their primary focus



Source: author's taxonomy and own elaboration based on desk research and analysis of listed frameworks.

The analysis of publicly available information on these frameworks revealed that they incorporate various ethical standards, but do not clarify who determines how these standards are applied; their optimal use is during an ex-post assessment of the impact of AI systems. Insufficient evidence supports the definition of a narrow implementation of responsible AI governance in the Public Sector. However, the analysis pointed out two relevant criteria to be of utmost importance for the definition of the author's proposed framework:

- A transdisciplinary approach that goes beyond involving public organisations, AI developers, and legal practitioners must incorporate the views of policy officials, regulatory bodies, and the public. All these stakeholders shall proactively ask themselves the above questions when using AI in algorithmic decision-making.
- There is a need to "Avoid the concentration of power" when distributing governance roles.

4.2 Definition of a multi-layered framework

When building the proposed framework, the author considers the complex dynamics ruling the cooperation among stakeholders and the flexibility of public-private AI appropriation schemes. Since building or refining a framework case by case is not possible or convenient, it became easier to adopt a qualitative model. The model aims to support the user, mainly the public authority or the policymaker, in detecting any activity that could lead to potential ethical warnings and assigning it to a procedural governance task. Moreover, it suggests focusing on those actions that can mitigate or eliminate the risk of ethical non-compliance in public processes and services when possible. Three aspects ground the choice for a qualitative typology:

- The ethics and responsibility principles are still at an abstraction level that makes it difficult, if not impossible, to measure their application or fulfilment.
- The fulfilment of ethics follows the progress of technology rather than measuring specific technological achievements or refinements.
- A qualitative model is more flexible than a quantitative one and better embeds the responsibility by design paradigm.

The approach of the author's study gathers these untold recommendations in the definition of a possible procedural model based on four fundamental pillars:

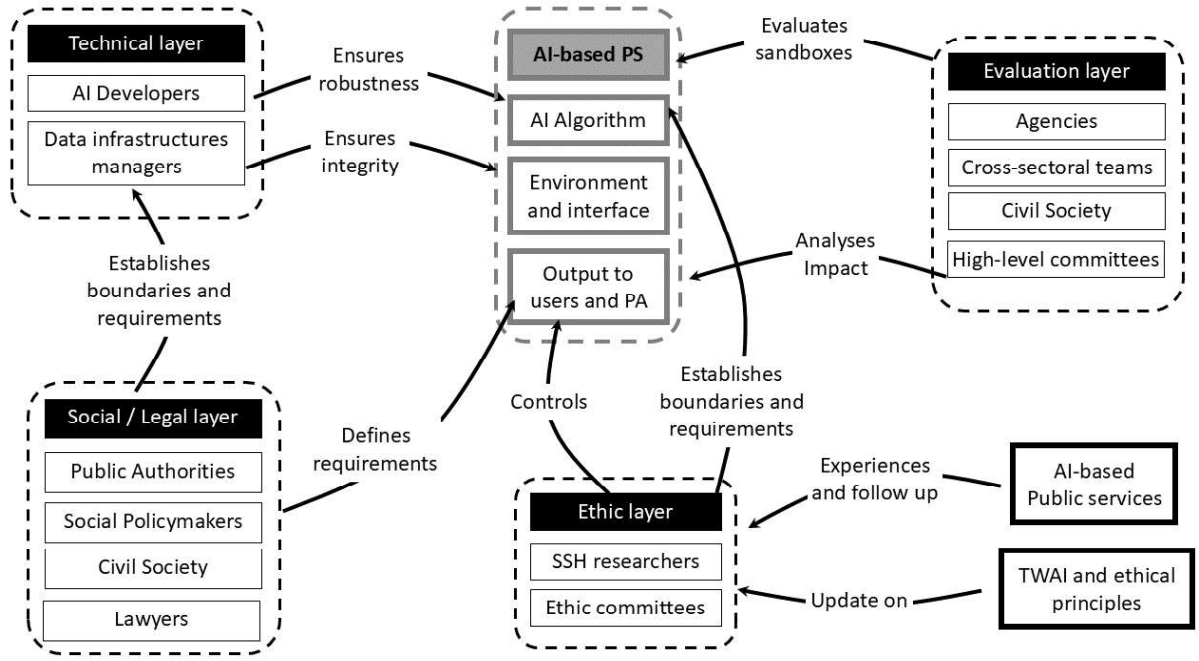
- Include a check on all criteria (Table 1) to evaluate any potential ethical breach.
- Consider the AI lifecycle as the sequential merging of three phases: AI design, AI appropriation, and AI (service) monitoring/evaluation.
- Require and check that all stakeholders are actively engaged

in the AI lifecycle, even if through different means.

- Identify rules and assign responsibilities to each stakeholder to ensure fair risk sharing and increase care for AI's overall quality.

Figure 2 introduces the schematic responsible human-centric AI governance model graphical representation. Each dashed area (governance layer) includes responsible and operative (who acts) stakeholders, while the arrows establish the connection on responsibilities (who sets the rules). For the sake of readability, the figure shows only the most relevant interdependencies among stakeholders of the various governance layers and the AI-based public services.

Figure 2: Schematic governance model representation



Source: Author's own elaboration, based on the match of Responsibility Models (Council of Europe, Yeung, K., 2019) and AI Governance Layers (Gasser, U., Almeida, V. A., 2017)

As said, responsible governance is the result of a refining exercise on each service and for each public authority providing the service. The customisation process could add or remove one or more stakeholders because of the technology's characteristics; accordingly, the scheme also includes double roles, especially in the evaluation layer. This layer oversees performing gap analysis in case some criteria are not adequately addressed and opts for a regulatory sandbox for the service's testing phase. On the other hand, the social and legal layers often decide which model of responsibility sharing to use, as they oversee the economic sustainability of the service and assign tasks to AI developers. The Ethics layer translates the key requirements into operational rules for the technical layer. Thanks to the experiential follow-up of other existing AI-based services, it provides enhanced control over the impact of the service on ethical principles. The approach of the governance model focuses on the assignment of clear roles and rules; therefore, when implementing the governance model, it is helpful to refer to a schematic procedure. In conclusion, the proposed governance model and the checklist method to implement it follow the strategies adopted by many of the other analysed frameworks. This study enhances this approach by adding three characteristics:

- The stakeholders' engagement ensures an adequate mapping of all the steps in the AI appropriation phase into the checklist with their responsibilities.
- The evaluation layer fulfils the need to constantly monitor the service's impact during its appropriation and operational phase.
- It embeds the "responsibility by design" paradigm, intending to support the public sector in operating ethically sustainable services from the very beginning and avoid ex-post legal monitoring by authorities.

Table 2 completes the governance model's scheme by regrouping all the missing details and specifying the tools and methods (third column). Although detailed, the table can be likewise refined. The nature of a governance tool enables service owners, providers, and public authorities or policymakers to fill all the table's cells with precise specifications and map roles for each engaged stakeholder.

Table 2 RHAIG framework detailed responsibility map

Layer	*Who "acts" and "is responsible for" in the layer*	*Who "sets the rules for" and "controls" the layer*	*With which "tools & methods"*
Technical	AI developers IT & Data Infrastructures managers IT and AI Research Centres	Public authorities National legal framework(s) Ethic committees (third parties)	Algorithm robustness check Ethic compliance (mainly security, bias, and exclusion)
Ethic	Ethic policy makers Social science researchers Ethic committees (third parties)	Ethic committees	the Implementation of TWAI principles Follow-up of AI services
Social / Legal	Public Authorities Social policy makers Civil society Lawyers	Public authorities Social policy makers	Definition of target territory and social boundaries Definition of models for allocation or responsibilities
Evaluation	Agencies (third parties) Cross-sectoral stakeholders' teams Civil Society High-level committees	Public authorities Legal framework	Constant KPI monitoring Sandboxes (and Living Labs) Citizens' feedback

Source: Author's own elaboration based on current research of mapped frameworks and responsibility models

The ex-post model's deployment highlights the RHAIG implementation in public AI-enabled services. Given the shortage of available data, the method relies on qualitative and Boolean analysis (YES / NO) to allow for a good evaluation rather than leveraging numerical indicators. From an operational perspective, the method requires checking which Trustworthy and Human Key Requirements are most relevant to the service and to what extent their governance has been fulfilled. The following checks are mandatory:

- Ensure to know (and have enough information about) which stakeholders are actively engaged in the whole AI-based service lifecycle: public authority adopting the service, AI developers, data infrastructure managers, social policy-makers, research centres, and – if possible – also some civil / citizen representatives.
- For each analysed criterion, all four layers of AI governance (technical, ethical, socio-legal, and evaluation) must be mapped as long as sufficient information about the existence and effectiveness of clear responsibilities among stakeholders for each layer is available.
- Discuss and analyse the possible consequences of the adopted AI based on all the criteria. The goal is to retrieve information on the Appropriate degree of human involvement, the risk of AI harm to individuals, the Plan for staff training (if necessary), and the Mechanisms for users' feedback.

5. A preliminary set of skills in Responsible Human-Centric AI Governance

Much work and research have been done to identify consolidated and emerging skills capable of AI implementation and

acquisition. Mindful of identifying the critical human-centric skills essential for working with and alongside AI, this study focuses on aligning the proposed framework model's needs with the existing peer-validated classification of mapped skills, rather than eliciting or defining new skills. Typical basic essential skills, such as Creativity, Critical Thinking, Collaboration, and Communication, can be developed «through education and training on the job» (Kumar, S., 2023). Human-centric AI, however, also requires understanding human behaviour and cognitive processes, human motivation, values, cognition, behaviour, learning, change, decision-making, and persuasion; human-centric AI design methods; Technology acceptance of interactive AI-infused systems; Design guidelines for Human-AI interaction; Algorithmic nudges and boosts, to name a few. The matching exercise goes in this direction by scouting additional skills through a cross-check between the skills/competencies classification and the requirements set by the proposed responsible AI framework, as mapped in Table 2.

The skills source for this study is a recent EC report, whose findings elaborate on the need to ensure the presence of proper competencies and governance practices to deploy AI solutions in the Public Sector. This approach facilitates the skills-matching exercise because it provides a global vision of the competencies needed to perform the governance practices effectively and assigns them to different responsibility and control layers, similar to our framework model approach. Notably, particular attention is given to structural governance practices, defined as «those that concern the identification of key decision-makers and their corresponding roles and responsibilities» (Medaglia, R., *et al.*, 2024), like the proposed framework. Moreover, its reliability leverages a synthesis of empirical research and grey and policy literature. The EC report defines three groups of

competencies: technical, managerial and policy, legal and ethical. For the sake of simplicity, the present study only considers the three groups; however, it is worth mentioning that the report also introduces three cross-cutting clusters drawing on the distinction between know-why, know-how and know-what types of knowledge. Matching the 56 different skills and competencies observed in the EC report with the requirements of the proposed framework brings the following as a starting base for the list of skills needed for ensuring the consistent implementation of responsible human-centric AI governance in the public sector:

1. Technical:

- Technical Design Thinking, *being able to approach innovation of AI technology in an iterative and user-centred way*
- Data quality assessment, *being able to assess the quality of data in AI contexts*
- Choice of machine learning techniques, *being able to know when to use a certain algorithm, tool or library in a specific situation*
- Compliance with AI standards, *being able to adhere to and develop AI based on ethical and legal technical standards*

2. Managerial:

- AI benefits understanding: *being able to understand the benefits of AI.*
- User centricity: *being able and willing to collaborate with users of AI digital services and valuing their feedback.*
- Risk anticipation and migration: *being able to anticipate and mitigate risks of AI (e.g. privacy, security and ethics).*
- Data-supported decision-making: *being able to make decisions based on data.*
- The choice to delegate to AI: *being able to consider relevant factors in deciding whether or not to delegate a public service or a process to AI.*

- Change management: *being able to manage changes in organisational processes introduced by AI.*

3. Policy-Legal-Ethical:

- Awareness of ethical implications: *being able to be aware of the implications of AI on ethical and moral issues.*
- Awareness of sustainability implications: *being able to be aware of the implications of AI on environmental sustainability*
- Policy Design thinking: *being able to approach AI policymaking in an iterative and user-centred way.*
- Collaboration with AI ethicists: *being able to judge when experts on AI ethics should be consulted.*
- Collaboration with domain experts: *being able to work together with domain experts from varying professional backgrounds.*
- Understanding legal and ethical frameworks: *being able to understand and be aware of relevant legal and ethical frameworks for AI.*
- Privacy and security literacy: *being able to understand and act on the issues, concerns and threats around privacy and security raised by AI.*

The list includes the minimum set of skills and competencies addressing responsible human-centric AI, whilst still not being exhaustive. Table 3 briefly addresses the skills mapping exercise by assigning the skills in the proposed scheme.

Table 3: Map of competencies and skills for the proposed RHAIG framework

Layer	*Who "acts" and "is responsible for" in the layer*	*Who "sets the rules for" and "controls" the layer*
Technical	Technical Design Thinking Data supported decision-making Collaboration with domain experts	Choice of machine learning techniques Collaboration with AI ethicists Change management
Ethic	Awareness of ethical implications Awareness of sustainability implications	Risk anticipation and migration
Social / Legal	Data quality assessment Understanding legal and ethical frameworks	Policy Design thinking Privacy and security literacy
Evaluation	Compliance with AI standards User centricity	Choice to delegate to AI AI benefits understanding

Source: Author's own elaboration, based on the proposed framework and the analysis of "Competences and governance practices for artificial intelligence in the public sector", 2024

Conclusion

This study provides a preliminary, methodologically grounded, albeit incomplete, answer to the central question of defining a model for Responsible Human-Centric AI Governance that incorporates ethics and methods defined by the European public sector's trustworthiness policies. The proposed model provides added value to the mapped existing frameworks by analysing the ethical principles that consider the inner (disaggregated) layer of criteria. To the author's knowledge, this detail is not considered now in the landscape of solutions for AI governance. It is noticed, however, that the OECD is currently developing a framework based on the Recommendation on AI (OECD, 2024) and the Tools for Trustworthy AI (OECD, 2021). The proposed structure is general-purpose and includes a checklist

covering various aspects, including the roles and duties of stakeholders. The model, therefore, requires information from multiple sources; extensive interviews or stakeholders' questionnaires optimise the quality, whereas desk-based exercises could limit its potential. The model's goal is to provide suggestions and recommendations on how risks of implementing uncompliant ethical AI in the Public Sector can be mitigated with procedures and by leveraging a minimum set of skills. The proposed model constitutes an attempt to guide public authorities willing to adopt AI-based solutions by suggesting how and with whom they engage and with which tools. Its procedures allow users to fill out the responsibility mapping table by spotting gaps identified by three parameters: WHAT (the criteria not adequately addressed and the possible action), WHO (the stakeholders mostly entitled to be responsible for that action), and the responsibility LAYER considered (if technical, ethical, legal, or evaluation). Another problem this study highlights is that, in the AI acquisition process, stakeholders often find ethical frameworks abstract and challenging to apply when developing or tuning systems, leaving unanswered questions, such as at what development stage an AI system should be tested against these principles. The proposed minimum set of skills serves as a starting point to address this knowledge and awareness gap. Yet, it is far from claiming to be a definitive solution for identifying the needs of AI practitioners in the context of human-centric AI. On the other hand, it shows that current mapped skills and competencies could play a relevant role in ensuring that ethics and trustworthiness are embedded in the human-centric AI deployment and acquisition by the Public Sector. Three main recommendations come from this study and represent interwoven issues:

- It is crucial to ensure the 'autonomy' of the responsibility assignment for any AI-based development. Establishing a

straightforward and flexible approach to jointly regulating AI seems relevant to ensure innovation while safeguarding fundamental values and preventing harm to people.

- Rules and regulations are useless without coherent training. One of the main challenges is motivating and engaging core stakeholders to ensure they act upon the given recommendations. Often, skills and knowledge constitute barriers and ensuring that all subjects are correctly trained on responsible procedures will smooth them out.

Investing in research on competencies for trustworthy, human-centric AI in the public sector should be a global policy. AI-based technologies continue to grow; generative AI applications are a prime example of how public servants require the latest competencies, rather than existing ones, to utilise large language model prompts for document drafting.

References

AI-HLEG (2019). High-Level Expert Group on Artificial Intelligence: Ethics guidelines for trustworthy AI.

Andrus, M., & Villeneuve, S. (2022, June). Demographic-reliant algorithmic fairness: Characterising the risks of demographic data collection in the pursuit of fairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1709-1721.

Council of Europe (2019). Responsibility and AI: Council of Europe Study DGI(2019)05.

European Commission (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act).

European Commission (2024a). Public Sector Tech Watch: Mapping innovation in the EU public services: a collective effort in exploring the applications of artificial intelligence and blockchain in the public sector.

European Commission (2024b). Public Sector Tech Watch: Adoption of AI, Blockchain and other emerging technologies within the European public sector. Publications Office of the European Union, Luxemburg, 2024.

Gasser, U., Almeida, V. A. (2017). A layered model for AI governance. *IEEE Internet Computing*, *21*(6), 58-62.

Grimmelikhuijsen, S. Tangi, L. (2024). *What factors influence perceived artificial intelligence adoption by public managers*. European Commission: Joint Research Centre. Luxembourg: Publications Office of the European Union. https://data.europa.eu/doi/10.2760/0179285, JRC138684.

Hak, A. (2022). Why AI governance is important for building more trustworthy, explainable AI. https://thenextweb.com/news/ai-governance-critical-trustworthy-explainable-ai (accessed on 18/10/2024).

Kumar, S. (2023). Developing Human Skills in the Era of Artificial Intelligence: Challenges and Opportunities for Education and Training. *Scholedge International Journal of Multidisciplinary &*

Allied Studies, *10*(2), 11-19. https://dx.doi.org/10.19085/sijmas100201.

Manzoni, M., *et al.* (2022). AI Watch. Road to the adoption of Artificial Intelligence by the public sector.

Medaglia, R., *et al.* (2024). Competences and governance practices for artificial intelligence in the public sector, Luxembourg: Publications Office of the European Union. European Commission: Joint Research Centre, https://data.europa.eu/doi/10.2760/7895569, JRC138702.

Misuraca, G. and Van Noordt, C. (2020). AI Watch - Artificial Intelligence in public services, EUR 30255 EN. Luxembourg: Publications Office of the European Union. doi:10.2760/039619, JRC120399. https://publications.jrc.ec.europa.eu/repository/handle/JRC126665.

Molinari, F., *et al.* (2021). AI Watch. Beyond pilots: sustainable implementation of AI in public services. EUR 30868 EN, Luxembourg: Publications Office of the European Union. doi:10.2760/440212, JRC126665.

OECD (2021), Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems. *OECD Digital Economy Papers*, *312*.

OECD (2024) Recommendation of the Council on Artificial Intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (accessed on 18/10/2024).

Pechtor, V., Basl, J. (2022). Analysis of suitable frameworks for artificial intelligence adoption in the public sector. IDIMT-2022.

Tangi L., *et al.*, (2022), AI Watch. European landscape on the Use of Artificial Intelligence by the Public Sector, EUR 31088 EN, Luxembourg: Publications Office of the European Union. doi:10.2760/39336, JRC129301.

UN (2024). Resolution on the promotion of safe, secure and trustworthy artificial intelligence systems for sustainable development. A/78/L.49.

Veenstra, A.F. *et al.* (2021). AI: in search of the human dimension. https://publications.tno.nl/publication/34638227/KTvx9Y/veenstra-2021-ai.pdf.

Yeung, K. (2019). *Responsibility and AI: A Study of the Implications of Advanced Digital Technologies (including AI Systems) for the Concept of Responsibility Within a Human Rights Framework*. https://rm.coe.int/a-study-of-the-implications-of-advanced-digital-technologies-including/168096bdab.

Zuiderwijk, A., *et al.* (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, *38*(3).