Publications Pertinentes

Contribution Récentes en Gouvernance et Gestion de l'Intelligence Artificielle

Carlos Denner dos Santos

Candidature — Poste de Professeure ou Professeur en Gestion de l'Intelligence Artificielle

Université de Sherbrooke, Département SIMQG

Novembre 2025

Publications Incluses

1. **Almeida, P. G. R. & Santos, C. D.** (2025). *Artificial Intelligence Governance : Understanding How Public Organizations Implement IT.* **Government Information Quarterly**, 42(1), 102003.

   Impact and Relevance : Publication la plus récente (2025). Étude empirique de la mise en œuvre de la gouvernance de l'IA dans 28 organisations publiques sur cinq continents. Directement alignée avec l'axe 1 du programme de recherche proposé.
2. **Almeida, P. G. R., Santos, C. D., & Farias, J. S.** (2021). *Artificial Intelligence Regulation : A Framework for Governance.* **Ethics and Information Technology**, 23(3), 505–525.

   Impact and Relevance : Publication fondatrice du programme de gouvernance de l'IA. Propose un cadre intégrateur pour la régulation de l'IA. Cite plus de 200 fois. Établit les concepts fondamentaux qui structurent le programme de recherche proposé à Sherbrooke.
3. **Moura, P. J., Santos, C. D., Bellini, C. G. P., & Dias, J. J. L.** (2024). *The Over-Concentration of Innovation and Firm-Specific Knowledge in the Artificial Intelligence Industry.* **Journal of the Knowledge Economy**, 15(4), 20547–20577.

   Impact and Relevance : Publication 2024 portant directement sur l'industrie de l'IA. Analyse la concentration de l'innovation et des connaissances spécifiques aux entreprises — enjeu clé de la durabilité et de l'impact des systèmes d'IA dans les organisations (Axe 3 du programme de recherche).

Note : Les trois publications ci-après sont présentées dans l'ordre de leur pertinence pour le poste. Elles couvrent collectivement : (1) la recherche empirique récente sur la gouvernance de l'IA dans les organisations, (2) le cadre conceptuel fondateur en régulation de l'IA, et (3) une preuve antérieure de rigueur en gouvernance organisationnelle complexe.

Government Information Quarterly 42 (2025) 102003

Contents lists available at ScienceDirect

Government Information Quarterly

journal homepage: www.elsevier.com/locate/govinf

Artificial intelligence governance: Understanding how public organizations implement it

Patricia Gomes Rêgo de Almeida[a,b,*], Carlos Denner dos Santos Júnior[a]

[a] University of Brasilia, Department of Business Management-PPGA, Campus Darcy Ribeiro, Prédio da FACE, Asa Norte, Brasília, DF 70910-900, Brazil
[b] Chamber of Deputies of Brazil, Palácio do Congresso Nacional - Praça dos Três Poderes, Brasília, DF 70160-900, Brazil

ARTICLE INFO

Keywords:
AI governance
AI regulation
Responsible AI
AI ethics
AI in public sector
AI in government
Fuzzy QCA

ABSTRACT

While observing the race for Artificial Intelligence (AI) regulation and global governance, public organizations are faced with the need to structure themselves so that their AI systems consider ethical principles. This research aimed to investigate how public organizations have incorporated the guidelines presented by academia, legislation, and international standards into their governance, management, and AI system development processes, focusing on ethical principles. Propositions were elaborated on the processes and practices recommended by literature specialized in AI governance. This entailed a comprehensive search that reached out to 28 public organizations across five continents that have AI systems in operation. Through an exploratory and descriptive aim, based on a qualitative and quantitative approach, the empirical analysis was carried out by means of proposition analysis using the Qualitative Comparative Analysis (QCA) method in crisp-set and fuzzy modes, based on questionnaire responses, combined with an interview and document content analysis. The analyses identified how processes and practices, across multiple layers and directed at the application of ethical principles in AI system production, have been combined and internalized in those public institutions. Organizations that trained decision-makers, AI system developers, and users showed a more advanced stage of AI governance; on the other hand, low scores were found on actions towards AI governance when those professionals did not receive any training. The results also revealed how governments can boost AI governance in public organizations by designing AI strategy, AI policy, AI ethical principles and publishing standards for that purpose to government agencies. The results also ground the design of the AIGov4Gov framework for public organizations to implement their own AI governance.

1. Introduction

After being coined “Artificial Intelligence” in 1956 (Cerka et al., 2017), a variety of research on AI has been developed, initially as knowledge-driven, later as data-driven, or combining them. The growing scope of AI in society has been observed through the combination of this technology with other emerging ones, generating the expression AI Plus (AI+) (Shao et al., 2022), boosting productivity (Mezgár & Váncza, 2022) and its influence on social transformation (Boyd & Holton, 2018).

The benefits offered by AI have reached the Government (Alhosani & Alhashmi, 2024). While Artificial Intelligence (AI) is seen as an enabler of digital transformation for organizations (Holmström, 2022; Kitsios & Kamariotou, 2021), in the public sector, governments' development strategies coincide with their AI strategies (Wirtz et al., 2018). However, at the same time, concerns grow about ethical impacts on AI-dependent decisions when ethical principles are not considered (Ashok et al., 2022; Bonsón et al., 2021; Hopster, 2021; Kazim & Koshiyama, 2021; Stahl et al., 2022; Wirtz et al., 2022). Immersed in this scenario, the movement for a responsible AI (Eke et al., 2023) using AI regulation and governance has involved governments, academia, and international standardization bodies (Carter, 2020; Gianni et al., 2022; Gutierrez & Marchant, 2021; IEEE, 2019, 2020, 2021a, 2021b, 2021c, 2021d, 2022; ISO, 2021a, 2021b, 2021c, 2022a, 2022b; OECD, 2022c).

Even though there is a significant number of theoretical essays (De Almeida et al., 2021), AI governance is still an underdeveloped area of research (Morley et al., 2020; Taeihagh, 2021), requiring a greater understanding of how organizations have interpreted and incorporated

[*] Corresponding author at: University of Brasilia, Department of Business Management, Campus Darcy Ribeiro, Prédio da FACE, Asa Norte, Brasília, DF 70910-900, Brazil.

E-mail addresses: patricia.rego.almeida@gmail.com (P.G.R. de Almeida), carlosdenner@unb.br (C.D. dos Santos Júnior).

https://doi.org/10.1016/j.giq.2024.102003
Received 11 June 2023; Received in revised form 25 September 2024; Accepted 20 December 2024
Available online 16 January 2025

ethical principles into their practices, processes and structures when producing AI systems (Mäntymäki et al., 2022; Mikalef, Conboy, et al., 2022). In systematic research on AI in public governance, Zuiderwijk et al. (2021) identified research gaps that adopt multiple methods, combining exploratory empirical research with qualitative-quantitative analyses, in order to delve deeper into practices used by AI governance in the public sector. Considering the important role that government bodies play in AI regulation and governance (Cihon et al., 2020; Stix, 2021), the potential benefits and risks that AI can bring to society when it supports the public sector (Ahn & Chen, 2022; Ojo et al., 2019; Sharma et al., 2020), and the challenge of avoiding loss of confidence in AI-supported government decisions (Zuiderwijk et al., 2021), it becomes crucial to understand how public organizations are following the academia, legislation, and standard recommendations concerning ethical principles in the use and development of AI systems.

2. AI governance

An AI that considers ethical principles brings new obligations to organizations (Hickman & Petrin, 2021; Roorda, 2021; Smuha, 2021). However, traditional governance processes and structures are not sufficient for the challenges posed by AI governance (Taeihagh, 2021). In public institutions, the challenge is greater because, in general, citizens do not choose AI products but are obligated to consume them as they are embedded in public services (Zuiderwijk et al., 2021).

In the context of AI governance in public organizations, ethical principles are applied to maximize the benefits of AI and minimize its risks (Rose et al., 2018; Vial, 2019). Considering the focus on impacts generated in society as a premise for AI governance (Djeffal, 2018), researchers point out the need to integrate the Stakeholder Theory with the Social Contract Theory (Bonsón et al., 2021; Wright & Schultz, 2018) to obtain society's perceptions of the values involved in decisions made by AI systems (Hickman & Petrin, 2021; Rahwan, 2017). In the corporate context, subordinated to IT governance (IT Governance Institute, 2003), an effective AI governance requires a multilayered model, the upper layer of which includes mechanisms for capturing government regulatory requirements and legislation, translating them into the organizational context through internal regulations that establish ethical principles and conditions for their application through processes and practices (Mäntymäki et al., 2022), as illustrated by the conceptual research model in Fig. 1a, whose details are found in the subsections 2.1 up to 2.6.

Considering many initiatives to regulate AI through legislation, government policies, and international standards (Fjeld et al., 2020; Gutierrez & Marchant, 2021; OECD, 2022c), it is expected that public organizations implement their own AI governance aligned with such regulation initiatives.

2.1. AI governance actions at the strategic level of public organizations

The relationship among the different scopes of governance, defined by Mäntymäki et al. (2022), establishes that corporate governance contains IT governance, which, in turn, contains AI governance. Consequently, AI governance inherits characteristics from IT governance.

IT governance relates to IT decision-making at the board of directors and executive management, which involves: creating an organizational structure (unit, committee, board), elaborating a strategy to effectively address the organization's needs through AI (Herremans, 2021), and implementing processes that support the board's decisions (Aasi et al., 2014). To formalize IT governance, organizations establish policies (Mäntymäki et al., 2022). Thus, similar to those actions that demonstrate the existence of IT governance at a higher-level decision-making board (Aasi et al., 2014), high-level decision-making actions for AI governance were considered in this research.

Indeed, in Papagiannidis et al.'s (2023) research, the higher-level decision-makers highlighted the importance of an AI strategy to manage the corporate needs that AI systems will address, as well as the AI governance process in their organization. In the same sense, Agarwal's (2023) research points out that the existence of an AI governance structure at a high level of the organization is crucial for setting the organization's strategic direction in AI-related initiatives. Complementing them, Sigfrids et al. (2023) emphasize that, in the public sector, AI policies should consider AI ethical principles, giving special

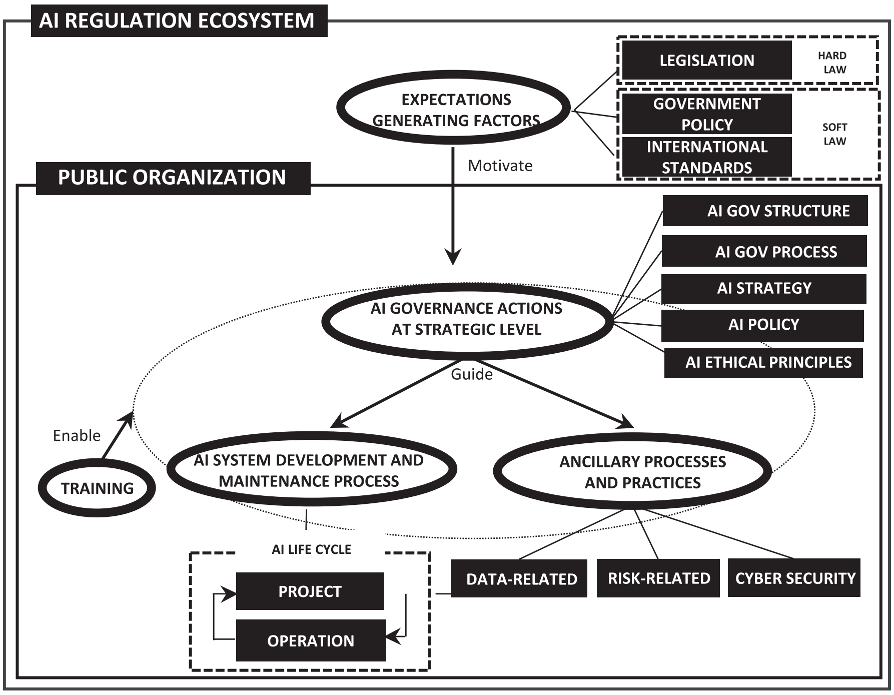


Fig. 1. a: Conceptual research model for AI governance in public organizations. (Source: Self-elaboration.)

attention to a wider socio-technical and political approach instead of merely respecting moral minimums. It requires considering values and gaining citizens' trust. Such an approach imposes establishing an AI ethical principles code at a high level of the organization. Thus, for the context of this research, "creating an AI governance structure," "elaborating an AI strategy," "establishing an AI policy," "implementing an AI governance process," and "establishing an AI ethical principles code" were at the higher level of the conceptual research model and were called "AI governance actions at a strategic level."

Guided by the AI governance actions at a strategic level, ancillary mechanisms are created to establish data-related processes (Janssen et al., 2020; Rhahla et al., 2021), a cybersecurity-related process (Breier et al., 2020; Xue et al., 2020), risk-related processes and practices (NIST, 2022; Wirtz et al., 2022) (Fig. 1a – Ancillary processes and practices) and AI system development-related processes that address AI ethical issues (De Silva & Alahakoon, 2022; Laato et al., 2022) (Fig. 1a – AI system development and maintenance process). Based on the above, the following **Proposition 1** is conjectured: "data-related," "risk-related," "cyber-security," and "AI system development" practices must follow guidelines established at the strategic level of AI governance. Considering that each dimension addressed in Proposition 1 can be decomposed into other actions, for greater accuracy of this study, derived propositions were created for each group of processes and practices.

2.2. Data-related processes and practices (Fig. 1a – Data-related processes and practices)

Given AI's reliance on data, a link is established between data quality and the outcome of AI systems (Dwivedi et al., 2021; Kuziemski & Misuraca, 2020). For this reason, focusing on reliable AI systems and data governance becomes crucial (Haneem et al., 2019; Vining et al., 2022).

Sometimes intersecting with AI governance (Alshahrani et al., 2021; Andrews, 2018; Dwivedi et al., 2021; Mäntymäki et al., 2022; Medaglia et al., 2021; Özdemir & Hekim, 2018), a data governance process specifies responsibilities over decisions made about the organization's data, as well as formalizes policies and standards (Abraham et al., 2019; Carretero et al., 2017; Vilminko-Heikkinen & Pekkola, 2019), which requires a great deal of stakeholders' negotiation skills to obtain consensus (Benfeldt et al., 2020; Calzada & Almirall, 2020; Micheli et al., 2020; Ruijer, 2021). Following data governance policies (Carretero et al., 2017), processes are defined for managing data quality (Haneem et al., 2019; Khatri, 2016) and personal data protection (Janssen et al., 2020). For those reasons, **Proposition 1 A** is based on the decomposition of the "data-related" construct into three processes: data governance process, data quality management process, and personal data protection management process, which must follow guidelines established at the strategic level of AI governance.

2.3. Risk-related processes and practices (Fig. 1a – risk-related processes and practices)

Proposition 1B comprises actions created to mitigate the risks posed by AI systems (Vetrò et al., 2021; Wirtz et al., 2022). However, traditional risk management processes, supported by quantitative methods (Chen & Deng, 2022; Duijm, 2015), have been criticized, requiring complementary approaches that integrate visions and thus also include a qualitative approach (Aiken, 2021; Fernandes et al., 2021; Gerkensmeier & Ratter, 2018; ISO, 2022a; ISO, 2022b).

Such integrated vision starts with the definition of stakeholders that are impacted, whether directly or indirectly, by the AI system (NIST, 2022; Wirtz et al., 2022). Therefore, identifying stakeholders in the whole AI lifecycle is a requirement for risk management (Wright & Schultz, 2018). In the same direction, audit processes for AI systems are also risk-oriented (De Oliveira, 2019; Erlina et al., 2020), associating them with stakeholders (Zicari et al., 2021). In addition, over time, changes in environmental variables can alter the context for which the AI system was designed, causing behaviors that differ from the desired results. This situation can be avoided by monitoring changes in the environment (rules, social trends, etc.) and feeding the risk management process (González et al., 2020). Thus, **Proposition 1B** was formulated as follows: risk management processes, audit processes, practices for identifying stakeholders, and practices for monitoring changes in the environment, all of which must follow guidelines established at the strategic level of AI governance.

2.4. Cybersecurity process (Fig. 1a – Cybersecurity)

Integrated with risk management (Breier et al., 2020; European Union Agency for Cyber Security, 2022), a cybersecurity management process is applied to prevent cyberattacks explicitly designed to exploit vulnerabilities in AI algorithms (Chen et al., 2019; Eggers & Sample, 2020; Gu et al., 2019; McGraw et al., 2020; Xue et al., 2020). To face those threats, organizations implement a security management process (European Union Agency for Cyber Security, 2021) that addresses the entire AI system lifecycle (Jing et al., 2021). Therefore, **Proposition 1C** was thus formulated as follows: the AI system security management process must follow guidelines established at the strategic level of AI governance.

2.5. AI system development process (Fig. 1a – AI system development and maintenance process)

Seeking to encompass the entire AI lifecycle, **Proposition 1D** was formulated as follows: practices aimed at ethical principles for AI system development and maintenance processes, both as a project and as an operational product, must follow guidelines established at the strategic level of AI governance. To investigate practices at each phase, Proposition 1D was further broken down into 1D1 and 1D2 to analyze the project phase and the operation phase, respectively.

At the project's starting point (Fig. 2 – Project), the translation of ethical principles into rules on the behavior of the system is deepened (Dennis et al., 2016; IEEE, 2021a), focusing on the definition of groups and attributes to be protected (González et al., 2020; ISO, 2021a; Rajkomar et al., 2018). Rules and ethical dilemmas are identified and analyzed (Anderson & Anderson, 2018; Awad et al., 2020; Bench-Capon & Modgil, 2017; Bonnemains et al., 2018; Locher & Bolander, 2019; Ma et al., 2018; Schrader & Ghosh, 2018; Zicari et al., 2021). Data extraction and preparation tasks demand attention to understand their characteristics and quality, preparing them for pre-processing (De Silva & Alahakoon, 2022). Techniques are applied to identify and minimize data biases (Ashokan & Haas, 2021; Baeza-Yates, 2018; ISO, 2021a; Leavy et al., 2020; Lin et al., 2021; Ntoutsi et al., 2020; Oneto & Chiappa, 2020; Roselli et al., 2019; Silberg & Manyika, 2019), while adjustments are made in the data sample (González et al., 2020). The building and validation of models involve algorithmic research, which requires decisions that also need to be free of bias (Abdollahi & Nasraoui, 2018; Ashokan & Haas, 2021; De Silva & Alahakoon, 2022; Makhlouf et al., 2021). Transparency practices are required to explain the system results (Adadi & Berrada, 2018; Arrieta et al., 2020; Das, 2020; Dazeley et al., 2021; Kale et al., 2022; Phillips et al., 2021). Focusing on the AI system project phase, **Proposition 1D1** was formulated: practices for representing rules and ethical dilemmas, practices for minimizing biases, and practices for providing transparency in the AI system development process must follow guidelines at the strategic level of AI governance.

The sensitivity to changes in the context for which the AI system was created, combined with the fact that AI models are less complex than social realities (Strauß, 2021), impose continuous monitoring after the AI system is in operation (Laato et al., 2022), which implies: a) automatic performance monitoring in charge of the IT staff (De Silva & Alahakoon, 2022; Fjeld et al., 2020; González et al., 2020), b) human oversight of the AI system behavior, usually by someone delegated by

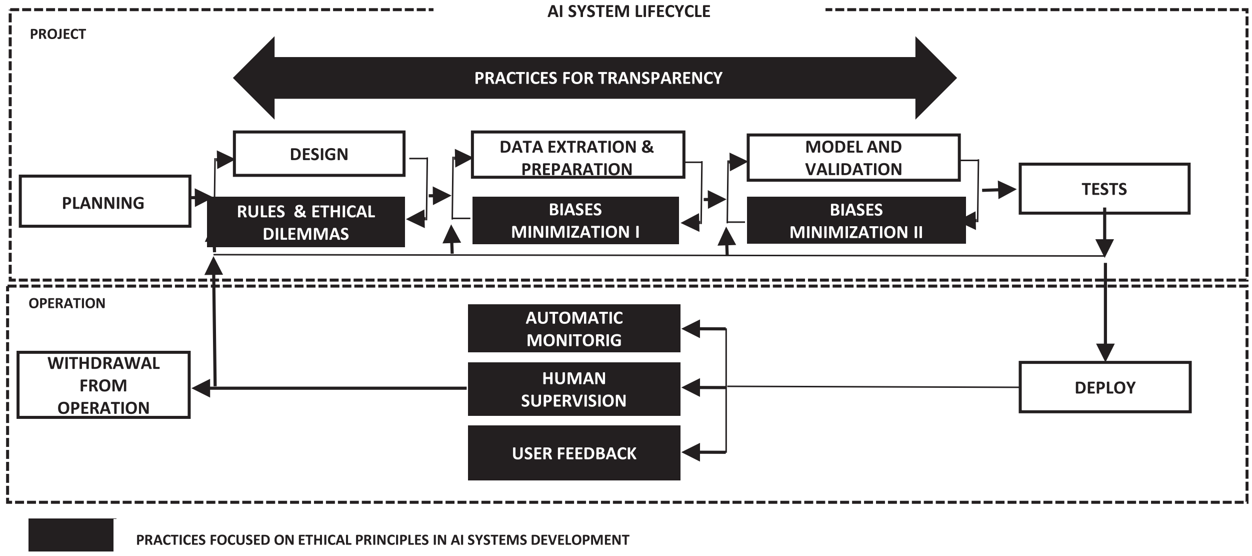


Fig. 2. Conceptual research model for the entire AI system lifecycle.
(Source: Self-elaboration)

the domains' decision-maker (Dignum, 2019; Fjeld et al., 2020; Hickman & Petrin, 2021; Strauß, 2021; Zicari et al., 2021), and c) continuous user feedback, which is often implemented through a feature in the AI system that asks its users' satisfaction with the system outputs and requests tips to refine its performance (Rahwan et al., 2019; Wright & Schultz, 2018) (Fig. 2 – Operation). The combination of those actions motivates the evolution of AI systems in a continuous loop until their withdrawal from operation (De Silva & Alahakoon, 2022; ISO, 2022a; Laato et al., 2022) (Fig. 2). Therefore, with the practices described in the AI system operation phase in mind, **Proposition 1D2** was formulated: practices for automatic monitoring, human oversight, and user feedback must follow guidelines established at the strategic level of AI governance.

2.6. *Training people (Fig. 1 – Training)*

Mikalef, Lemmer, et al. (2022) confirm that the decision-makers' perceptions regarding the potential AI value are drivers for AI adoption in public organizations. In the same sense, educating stakeholders on AI ethics becomes crucial for an effective AI ethical principles implementation in AI system production (Zhou & Chen, 2023), which includes a culture with fewer biases in organizations (Awad et al., 2020; Ma et al., 2018) as well as knowledge about practices required to AI governance (Ligot, 2024). With such purposes, training key stakeholders (decision-makers, developers, system users, and auditors) on AI ethical principles is considered an enabler for AI governance implementation in organizations (Calzada & Almirall, 2020; Herremans, 2021; Makarius et al., 2020; Micheli et al., 2020; Ruijer, 2021). Researchers point to an AI literacy program as a pivotal action to an effective and responsible AI adoption since it is a set of training and awareness actions that reaches staff beyond the IT team, which includes business decision-makers and system users. Decision-makers need it for a wide understanding of how AI can create value for their work processes and the requirements necessary for it (data quality, data protection, data governance, inclusive teams, for example). Users are also considered key stakeholders due to their role in data quality and data protection processes, and there is also the need for an awareness approach regarding ethical implications when AI is used inadequately (Pinski et al., 2024; Schüller, 2022).

Aiming to determine the enabling requirements of training people for AI governance practices, **Proposition 2** was elaborated: training stakeholders on data, the development of AI systems, and ethical principles applied to AI enables the implementation of AI governance in public organizations.

Considering that the complexity of internalizing AI governance in the organizations' processes and structures (Agarwal, 2023) can impact the "make-or-buy" decision on developing AI systems. Gräf et al. (2024) highlight that the opacity of AI systems and the lack of skilled human resources to deal with the whole AI system lifecycle are key factors for such decisions. In general, the decision to buy AI systems is an obstacle to accessing their codes, resulting in a transparency issue (Martin & Parmar, 2024). Adding such concerns to Mikalef, Lemmer, et al. (2022), Ahn and Chen's (2022) and Benfeldt et al.'s (2020) recommendations for training both the managerial and the technical spheres to deal with challenges such as reducing algorithmic opacity, it is possible to formulate **Proposition 3**: over time, in public organizations, training managers and AI system developers on AI ethics sparks interest in obtaining knowledge of AI system codes and contributes positively to AI governance.

3. Research methods and techniques

As an exploratory and descriptive study, the investigation was carried out through empirical research with a qualitative and quantitative approach to fill the gap identified by Mäntymäki et al. (2022) and Zuiderwijk et al. (2021) concerning knowledge of how organizations have interpreted and incorporated ethical principles in AI system production into their practices and processes, as especially demanded by Zuiderwijk et al. (2021) for using data-driven methods with exploratory and multiple approaches to deepen AI governance in the public sector.

3.1. *Sample selection and data collection strategy*

Since AI governance is a global need (Fjeld et al., 2020; OECD, 2022a), an attempt was made to build the sample including populations of public organizations belonging to any branches — Executive, Legislative, or Judiciary (Maluf, 1995) — from five continents. Considering the interest in investigating processes and practices, a search criterion was defined: a public organization should have at least one AI system in operation as part of its official portfolio. As a consequence, organizations that only had AI systems at a prototype stage or projects in an embryonic stage were not included in the sample.

The identification, sample selection, and data collection tasks were performed over three months, consisting of several steps in parallel lines of research involving different actors and sources of information, as shown in Fig. 3 (MRE, 2022; OECD, 2022a); European Commission, 2018). Those communications took place through remote meetings, email, and social media until the contact information from researchers, government and AI system development and research centers, managers in charge of AI or digital transformation strategies, and AI use cases could be obtained. The path described in Fig. 3 culminated in 711 AI systems being developed or used by public organizations (European Commission, 2021b; FCT, 2021; Government of India, 2020; IPS-X, 2021; Misuraca & van Noordt, 2020; OECD/CAF, 2022; Tangi et al., 2022; WEF, 2020, 2020b). After removing redundancies and non-operational systems (prototypes or withdrawn from use), finetuning continued until the contact information for organizations that met the search criteria could be found. Upon invitation, 28 organizations effectively participated in the survey.

3.2. Data collection mechanisms

For the quantitative and qualitative analyses, primary data was used by means of an online questionnaire and semi-structured interviews (Appendix A - Annex 1a and 1b), both of which required basic knowledge of the decisions, processes, and practices related to AI system production. For this reason, the questionnaire was sent out only after the organization indicated a person in charge of the AI system portfolio. The interviews were conducted remotely and, in some cases, two or three individuals represented the organization.

Both the questionnaire and the interview script were submitted to a group of judges for assessment (PhD and Master researchers on ethical AI and data analysis) using methods indicated in the literature for each case (Fig. 4). For the questionnaire, the “Content Validation Coefficient” (CVC) (Aburachid & Greco, 2011; Hernández-Nieto, 2002; Silveira et al., 2018) was used with each question being evaluated on a scale from 1 to 5 to measure levels of clarity and relevance for the research (Appendix A - Annex 2 A). The interview script was evaluated using the “Validation for Qualitative Research Instruments” method (VALI-QUALI) (Torlig et al., 2022), an evolution of MRPQ (Torlig et al., 2019), taking the “Content” and “Semantics” dimensions into consideration. The content evaluation provided a score from 1 to 5 for each question in relation to the “alignment of each question with the research objective” and “adherence of the question to the investigated construct” attributes. The semantic analysis considered the “clarity” and “qualitative expectation of answer for each question” attributes (Appendix A - Annex 2). Pre-tests were carried out to the complete set (questionnaire and interview) using people with a similar profile to the target audience to confirm alignment with the research objectives (Manzini, 2004).

For each question regarding the existence of a practice or process implemented (or being implemented), four scores were considered according to standard answers (100 for “Yes, completely,” 67 for “Yes, but only partially,” 33 for “No, but a formal decision has been made to implement it,” and 0 for “No, and no formal decision has been made to implement it.” For the AI governance process and for practices aimed at transparency, adapted answers were made, once they were gathered from the interview (Appendix A - Annex 3).

3.3. Analysis strategy

The analysis of the primary data was carried out using a combination of Qualitative Comparative Analysis - QCA (Ragin, 2008; Rihoux & Ragin, 2008), and content analysis of the interviews and shared documents (Krippendorff, 2013; Saldaña, 2013).

3.3.1. QCA

The QCA is a qualitative research technique that also considers quantitative aspects for samples from 3 to 250 cases (Dias, 2011). Based on Set Theory and Boolean operations to establish logical relationships between sets, the QCA proposes to solve problems whose analysis requires causal inferences in case studies. The method seeks to show which combinations of conditions occurred in a scenario of an expected outcome (Rihoux & Ragin, 2008) to carry out comparative analyses, through associations between certain conditions and the outcome, instead of correlations (Korjani & Mendel, 2012; Ragin, 2008). A condition is required for a given result if the condition is always present when the outcome occurs. A condition is sufficient for a certain outcome

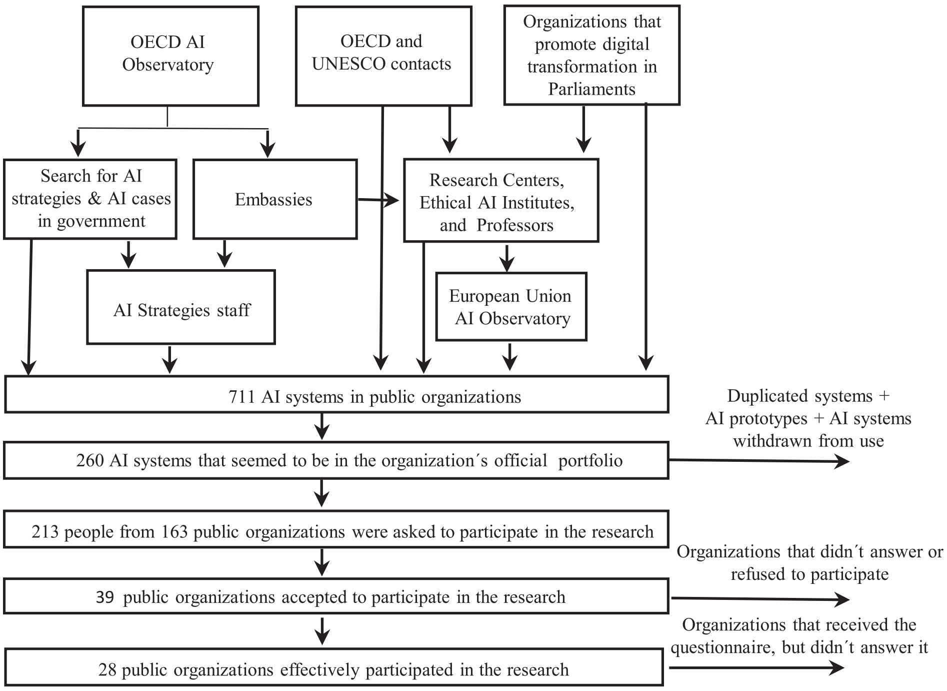

OECD AI Observatory
OECD and UNESCO contacts
Organizations that promote digital transformation in Parliaments
Search for AI strategies & AI cases in government
Embassies
Research Centers, Ethical AI Institutes, and Professors
AI Strategies staff
European Union AI Observatory
711 AI systems in public organizations
Duplicated systems + AI prototypes + AI systems withdrawn from use
260 AI systems that seemed to be in the organization´s official portfolio
213 people from 163 public organizations were asked to participate in the research
Organizations that didn´t answer or refused to participate
39 public organizations accepted to participate in the research
Organizations that received the questionnaire, but didn´t answer it
28 public organizations effectively participated in the research

Fig. 3. Sample selection strategy.
(Source: Self elaboration)

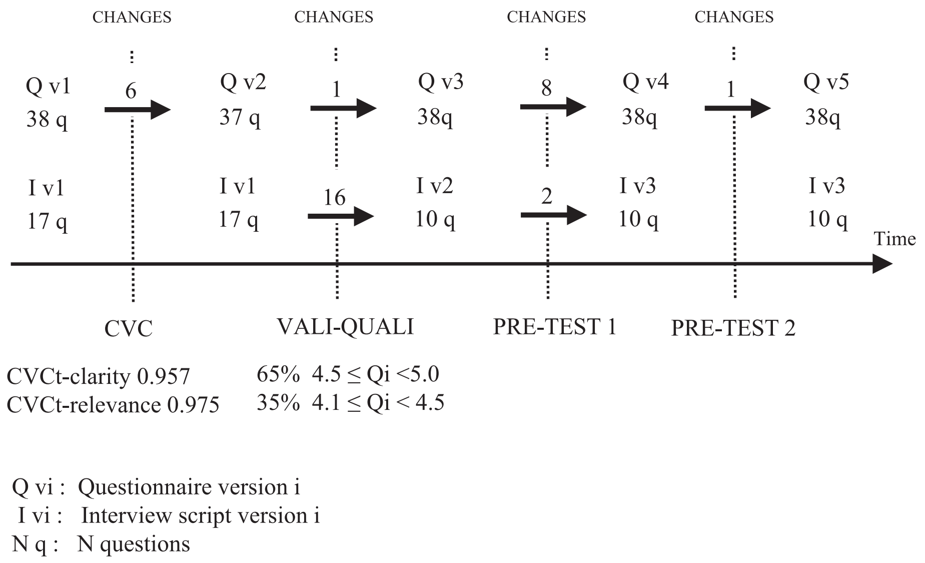


Fig. 4. Evolution of data gathering mechanisms.
(Source: Self-elaboration)

if this result always occurs when the condition is present (Rihoux & Ragin, 2008).

This research used both the crisp-set QCA (values 0 and 1, respectively, for the absence or presence of a relationship between the sets) and the fuzzy QCA, which offers more precision due to the use of a continuous set of values in the interval from 0 (complete absence of membership) to 1 (full membership) (Ragin, 2008; Rihoux & Ragin, 2008). In fuzzy analyses, A is a fuzzy subset of a fuzzy set B if given two fuzzy sets, A = {s1,s2,.., sn} and B = {g1,g2,.., gn}, and si, and gi, being i-th case scores in each set, and si ϵ [0;1] ⊂ Ʀ, gi ϵ [0;1] ⊂ Ʀ, ∀i; so, A ⊂ B if si ≤ gi, ∀i. A calibration (Freitas & Neto, 2016; Meijerink & Bondarouk, 2018; Ragin, 2008) was made using the original values of the sets, turning them into fuzzy sets, whose values are distributed in the interval between 0 and 1, based on the presence level of the conditions in the outcome set. The main criteria for validating the fuzzy QCA is the consistency indicator to measure the relationship proximity between sets, indicating the degree to which the cases that share a combination of conditions agree with the outcome, with a value between 0 and 1. Consistency $(Xi \leq Yi) = \sum \frac{MIN(Xi,Yi)}{\sum Xi}, \forall i$; where X is the membership score in the causal combination, and Y is the membership score in the outcome (Betarelli-Júnior & Ferreira, 2018). Complementing the outcome interpretation, the coverage indicator offers the quantification of the empirical relevance of a causal combination in the causal combination set: Coverage $(Xi \leq Yi) = \sum \frac{MIN(Xi,Yi)}{\sum Yi}, \forall i$; where X is the membership score in the causal combination and Y is the membership score in the outcome (Rihoux & Ragin, 2008).

3.3.2. *Analysis plan*

A crisp-set QCA was applied for dichotomous variables, and the fuzzy QCA for variables with continuous values (Fig. 5) was associated with the research model constructs. The QCA was applied to analyze the propositions. As for the crisp-set QCA, the TOSMANA software version 1.6.1 (https://www.tosmana.net/) was used, while fsQCA version 3.1 (http://www.socsci.uci.edu/~cragin/fsQCA/software.shtml) was utilized for the fuzzy QCA. Another piece of information was extracted from the analysis of the interview content using MAXQDA 2022 software (https://www.maxqda.com/). Lastly, the union of the two analyses grounded the discussions to reach the conclusions that support the proposed framework.

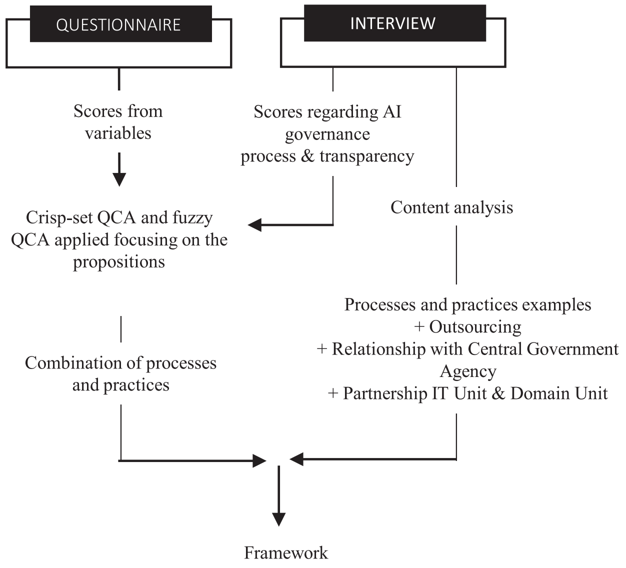

QUESTIONNAIRE
INTERVIEW
Scores from variables
Scores regarding AI governance process & transparency
Content analysis
Crisp-set QCA and fuzzy QCA applied focusing on the propositions
Processes and practices examples
+ Outsourcing
+ Relationship with Central Government Agency
+ Partnership IT Unit & Domain Unit
Combination of processes and practices
Framework

Fig. 5. Analysis plan.
(Source: Self-elaboration)

4. Results and discussions

The participants were people aware of the processes involved in the whole AI lifecycle. They were decision-makers and/or people appointed by them to participate. Table 1a displays the participants' profiles. The sample, resulting from the path taken as shown in Fig. 3, consists of 28 public organizations distributed across five continents, whose characteristics are listed in Tables 1b and 1c. The category of public organizations reveals their importance to citizens. The questionnaire asked the participant to provide information about the organization's five most frequently used AI systems (Appendix A - Annex 5).

The analyses that simultaneously involved dichotomous variables and continuous-value variables were performed by converting variables from continuous into binary and applying the crisp-set QCA (Proposition 3). For analyses in which all variables had continuous values (Propositions 1 A, 1B, 1C, 1D1, 1D2, and 2), the fuzzy QCA was applied with a fuzzy value calibration of 0.05 and 0.95, respectively, for each set's lowest and highest values, as also established by Navarro et al. (2016)

Table 1a
Interviewed profiles.

Organization Unit	N	%	Education	N	%	Position	N	%
Information Technology	17	60.7	PhD	9	32.1	Director/Manager	20	71.4
Innovation	2	7.14	Master	12	42.9	Data Scientist/IT Analyst	6	21.4
Data Science and Statistics*	3	10.7	MBA	4	14.3	Consultant	2	7.14
Corporate Governance	4	14.3	Undergraduate	3	10.7			
Others	2	7.14						

* Cases in which the data science unit is detached from the IT unit.

Tables 1b and 1c
Characteristics of the 28 public organizations in the sample used in the study.

Country	N	%
Angola	1	3.57
Argentina	3	10.71
Australia	1	3.57
Brazil	4	14.29
Canada	2	7.14
Denmark	1	3.57
Estonia	2	7.14
Finland	2	7.14
Germany	2	7.14
Iceland	1	3.57
Italy	1	3.57
Japan	1	3.57
Luxembourg	1	3.57
Norway	3	10.71
Sweden	1	3.57
Switzerland	1	3.57

Coverage	N	%
National	24	85.71
Group of countries	1	3.57
State/City	3	10.71

Branch of Government	**N**	**%**
Executive	17	60.7
Legislative	9	32.1
Judiciary	2	7.14

No of Employees	**N**	**%**
Up to 100	2	7.14
101 up to 500	3	10.71
501 up to 1000	8	28.57
1001 up to 10,000	12	42.86
More than 10,000	3	10.71

Category	N	%
Parliament	8	28.57
Ministry of Finances\National Agency for Trade and Investment \National Business Authority \National Tax Agency\National Agency for Improving Business	6	21.42
National Agency for Unemployment Insurance Fund\Federal Office for Migration and Refugees\National Agency for Labour and Welfare	3	10.71
National Statistic and Research Institute\National Institute of Research and Technology	3	10.71
Federal Court/Federal Court Dept Specialized in Child Sexual Abuse Crimes	2	7.14
Hospital and Health Research Institute/National Agency for Auditing Businesses Producing Food	2	7.14
National Agency for Account Auditing	1	3.57
State Agency for Innovation	1	3.57
State Agency for Public Transport	1	3.57
Chief of Ministerial President Cabinet	1	3.57

AI systems production time	**N**	**%**
Less than 1 year	1	3.57
1–3 years	8	28.57
3–5 years	11	39.29
More than 5 years	8	28.57

and Codá et al. (2022). The descriptions of the variables are in Appendix A - Annex 6.

Regarding legislation, it is worth mentioning that 100 % of the sample comprises organizations subject to some law that protects personal data. In the sample's scope, only Denmark (Danish Government, 2020) approved a law on data ethics that also addresses AI systems development. Among the bills under discussion, the European AI Act (European Commission, 2021a; OECD, 2022b), despite being in an advanced discussion stage when the data were collected, was not yet approved as a law (European Parliament, 2022a; European Parliament, 2022b).

4.1. Analysis of proposition 1

Proposition 1 was analyzed through its decomposition into dimensions – data, risks, cyber security, and development – in Propositions 1 A, 1B, 1C, and 1D analyses, respectively. In those analyses, to represent AI governance actions at the strategic level of the organization, the following was considered (Appendix A - Annex 4): AI strategy; policy or recommendations directed at AI systems; guide to ethical principles that are applied to AI systems; AI governance processes; and the existence of a structure (unit, position, or board/council/committee to address AI governance). Those analyses carried out the calibration for turning the original values into fuzzy values while considering that the analyses focus on the conditions whose practices and processes have been implemented in any proportion (scoring options 67 or 100 – Appendix A - Annex 3). For those purposes, in all research analyses that used a fuzzy QCA, the crossing point was defined as 60.

4.1.1. Analysis of proposition 1 A

From the decomposition of the practices regarding data into processes for data governance (PDGOV), for data quality management (PDQUA), and for personal data protection management (PDPRO), an analysis of Proposition 1 A was carried out, considering, as an output set, the cases that were in a more advanced stage of implementing AI governance actions at a strategic level (SAIGOV(1)) (Appendix A - Annex 6). The process for managing personal data protection had the highest average (90.5, see Appendix A - Annex 4), which illustrates the most advanced stage of actions to comply with the personal data protection law to which they are subject.

The three preliminary combinations presented by the fuzzy QCA (Table 2) were not valid to the conclusion (combinations 1 and 2 had a consistency below the minimum value accepted by Ragin, 2008 and Scheider and Wagemann, 2012, and combination 3 was not conclusive). However, it should be noted that nineteen cases implemented the three processes – data governance, data quality management, and personal data protection management; and that, of the twelve highest scores for strategic actions towards AI governance, nine cases (75 %, see Appendix A - Annex 4) had implemented, at any stage, all three processes. Due to that, the "Necessary Condition" assessment test was applied to the combination of the three processes, which resulted in a consistency of 0.959862, revealing that high values would have been unlikely to occur in AI governance actions at a strategic level without the simultaneous existence of the three processes tested. Thus, the simultaneous implementation of the three processes is associated with organizations at an advanced stage in those AI governance actions at the strategic level defined in 2.1.

The positive impact when the three data processes are implemented aligns with the researchers' arguments that data governance is crucial for AI governance (Haneem et al., 2019; Vining et al., 2022) since data governance defines stakeholders' roles and responsibilities concerning the data (Sivarajah et al., 2017), which, in turn, requires managers' involvement in many deliberations (Benfeldt et al., 2020) to improve data quality process (Haneem et al., 2019; Rhahla et al., 2021; Vilminko-Heikkinen & Pekkola, 2019) and personal data protection management process. Thus, data governance guides data quality management processes and personal data protection processes (Labadie et al., 2020).

Table 2
Fuzzy QCA applied to proposition 1 A.

Proposition		Combinations	Cases	Results	Coverage and Consistency
1A	1	PDGOV * PDPRO	3,5,9,19,22,27,6,8,13,25	SAIGOV(1)	Cv: 0.825606 Cs: 0.71781
			4,20,24,2,7,11,15,16,23	SAIGOV(0)	
	2	PDQUA * PDPRO	8,19,22,25,5,6,9,13,27	SAIGOV(1)	Cv: 0.733564 Cs: 0.688312
			20,23,2,4,7,21,15,16,18,28	SAIGOV(0)	
	3	~ PDGOV* ~ PDQUA* ~ PDPRO	21	SAIGOV(1)	Cv: 0.215917 Cs: 0.75
			1	SAIGOV(0)	
	"Necessary Conditions" Test	PDGOV * PDQUA * PDPRO		SAIGOV(1)	Cv:0.581795 Cs: 0.959862

4.1.2. *Analysis of proposition 1B*

Proposition 1B was analyzed based on the decomposition of risk-related practices into a process for risk management (PRISM), a process for auditing AI systems (PAUDIT), a practice for identifying stakeholders (STAKEH), monitoring changes in the environment and social trends (ENVIRO), considering cases that were at a more advanced stage in the implementation of AI governance actions at a strategic level (SAIGOV(1)) (Appendix A - Annex 6). Of the three combinations that were calculated (Table 3), only combination 2 showed a proportion favorable to advanced-stage cases in AI governance actions at a strategic level, which presented 85.71 % of the cases (cases 3, 8, 9, 13, 22, 27) that have implemented, at any stage, a risk management process, practices for identifying stakeholders, and practices for monitoring environmental changes and social trends (consistency = 0.941645). Therefore, those processes and practices can be considered to have been associated with more advanced stages in implementing AI governance actions at a strategic level.

The positive impact of the four studied practices aligns with the researchers' arguments that a risk management approach for AI requires that traditional risk management processes (Chen & Deng, 2022; Duijm, 2015) be associated with stakeholders' identification (Schaefer et al., 2021; Wirtz et al., 2022; Wright & Schultz, 2018), with a risk-oriented audit process (De Oliveira, 2019; Erlina et al., 2020), and with a change management in the environment (González et al., 2020).

4.1.3. *Analysis of proposition 1C*

Considering only the cyber security management process, this analysis did not present combinations of processes or practices because this process was not decomposed into other practices. The fuzzy QCA analysis (Table 4) revealed that, of the cases that had implemented a process for managing security in AI systems (PSEC), 66.67 % (cases 3, 5, 6, 8, 9, 13, 19, 22, 25, 26) obtained a high score for AI governance actions at a strategic level (SAIGOV(1)) (Appendix A - Annex 6), which indicates an association between the existence of a security management process and advanced stages of AI governance actions at a strategic level. Such result confirms researchers' arguments that a) AI governance implies, among other principles, ensuring robust and safe AI systems (Dalrymple et al., 2024), b) since robust and safe AI systems require practices to deal with mechanisms to face cyberattacks created specifically to AI systems (Booth et al., 2023; Ee et al., 2024), AI governance also requires practices systematically organized to manage the cyber security of AI systems. Compared with the other propositions' analyses, this was the lower consistency (0.8312) in the considered associations, which can be further investigated in future research.

4.1.4. *Analysis of proposition 1D*

The analysis of system development practices required two levels of decomposition: firstly, the phases of the AI system development and support process – project (Proposition 1D1) and operation (1D2); and subsequently, decomposing each of those phases.

4.1.4.1. Analysis of Proposition 1D1. The fuzzy QCA applied to the project phase practices (Table 5) used the representation of rules and ethical dilemmas (RDILEM), practices to minimize biases (PBIAS), and practices to provide transparency (TRANSP) in the AI system development (Appendix A - Annex 6). The low average score for practices aimed at transparency (50.11, see Appendix A - Annex 4) confirms the challenge of obtaining an explanation for the algorithm's results (Buiten, 2019; Butterworth, 2018; Zuiderwijk et al., 2021). Among other factors, this scenario may have been amplified by the fact that 73.33 % of the sample outsourced at least part of their AI system development, and only 46.43 % of the sample had access to their AI system code. The low average score for the representation of principles and ethical dilemmas (51.11, see Appendix A - Annex 4) may be due to the lack of a clear definition of those principles or the lack of specialists to implement the practice, as Ahn and Chen (2022) have alerted. Among the cases that had deployed practices to represent the business rules, principles and ethical dilemmas, and practices for transparency in AI system development, 75 % (cases 3, 9, 13, 21, 26, 27) showed high scores for AI governance actions at a strategic level (SAIGOV(1)).

When going deeper into the fuzzy-value analysis, we notice that those cases also implemented practices to minimize biases. The "necessary condition" test showed greater consistency (Cs = 0.96263) for the combination of the three practices. Therefore, the analysis of Proposition 1D1 revealed that in the studied sample, organizations that implemented practices to represent rules related to principles and ethical dilemmas, practices to provide transparency, and practices to minimize biases in AI systems during their development are associated with a more advanced stage in the implementation of AI governance actions at a strategic level.

Such result aligns with the researchers' arguments that, to ensure trustworthy AI systems, it is necessary to implement practices for

Table 3
Fuzzy QCA applied to Proposition 1B.

Proposition	Combinations	Cases	Results	Coverage and Consistency

1B	1. ~PAUDIT * STAKEH * ENVIRO	27,25
		2,16
			SAIGOV(1)
			SAIGOV(0)
				Cv:0.319031
				Cs:0.893411

	2. PRISM * STAKEH * ENVIRO	3,9,13,22,27,8
		2
			SAIGOV(1)
			SAIGOV(0)
				Cv: 0.49135
				Cs: 0.941645

	3.~PRISM*PAUDIT*STAKEH* ~ ENVIRO	5,21	SAIGOV(1)	Cv: 0.293426 Cs: 0.925764

Table 4
Fuzzy QCA applied to Proposition 1C.

Proposition	Combinations	Cases	Results	Coverage and Consistency

1C	PSEC	5,8,9,13,22,3,6,19,25,26
		11,2,16,20,28
			SAIGOV(1)
			SAIGOV(0)
				Cv:0.719031
				Cs: 0.8312

Table 5
Fuzzy QCA applied to Proposition 1D1.

Proposition	Combinations		Cases	Results	Coverage and Consistency

	RDILEM * TRANSP		9,13,27,3,21,26
			15,24
				SAIGOV(1)
				SAIGOV(0)
					Cv: 0.538408
					Cs: 0.913146

1D1	"Necessary Conditions" Test	RDILEM * TRANSP		SAIGOV(1)	Cv: 0.727125 Cs: 0.929412
		RDILEM * TRANSP * PBIAS		SAIGOV(1)	Cv: 0.679531 Cs: 0.96263

Table 6
Fuzzy QCA applied to Proposition 1D2.

Proposition	Combinations		Cases	Results	Coverage and Consistency

1D2	HOVER * FEEDB		3,6,9,13,22,25,5,14,
			19,26,27
			10,24,16,20,28
				SAIGOV(1)
				
				SAIGOV(0)
					Cv: 0.741869
					Cs: 0.790561

	"Necessary Conditions" Test	HOVER* FEEDB			Cv: 0.627069 Cs: 0.891349
		AMONI*HOVER*FEEDB			Cv: 0.576529 Cs: 0.972318

improving transparency along with the AI system development (Adadi & Berrada, 2018; Arrieta et al., 2020; Das, 2020; Dazeley et al., 2021; Kale et al., 2022; Phillips et al., 2021; Schaefer et al., 2021), for identification of ethical principles and dilemmas to be applied in the business rules (González et al., 2020; Rajkomar et al., 2018), and for mitigating biases in data preparation and in modeling (Ashokan & Haas, 2021; Baeza-Yates, 2018; De Silva & Alahakoon, 2022; Leavy et al., 2020; Lin et al., 2021; Makhlouf et al., 2021; Ntoutsi et al., 2020; Oneto & Chiappa, 2020; Silberg & Manyika, 2019).

4.1.4.2. Analysis of proposition 1D2. The analysis of practices in the operation phase of the AI system development and maintenance process (Table 6) considered the automatic monitoring (AMONI), human oversight (HOVER), and user feedback (FEEDB) variables (Appendix A - Annex 6). The highest average score was identified among the practices of this phase in automatic monitoring (71.57, see Appendix A - Annex 4), revealing the greater ease of monitoring when one does not depend on human resources. The preliminary fuzzy QCA showed a combination composed of human oversight and user feedback in sixteen organizations, of which 68.75 % (cases 3, 5, 6, 9, 13, 14, 19, 22, 25, 26, 27) obtained high scores for the AI governance actions at a strategic level (SAIGOV(1))). It is important to note that fifteen cases implemented automatic monitoring, human oversight, and user feedback, confirming the perceptions of Rahwan et al. (2019), Wright and Schultz (2018), and De Silva and Alahakoon (2022). Additionally, the "Necessary Conditions" test indicates that a high score would unlikely be obtained for AI governance actions at a strategic level without implementing the three practices (consistency = 0.972318). Therefore, there is an association between organizations at a more advanced stage in AI governance actions at a strategic level and cases that had automatic monitoring, human oversight, and user feedback practices, as argued by De Silva and Alahakoon (2022), Laato et al. (2022), Strauß (2021), González et al. (2020), and Zicari et al. (2021), when they demanded monitoring AI in the real environment.

4.2. Analysis of proposition 2

The analysis of Proposition 2 (Table 7) included the fuzzy QCA variables (Appendix A - Annex 6) training in data, AI risks, and AI ethical principles directed at decision-makers (TDEMAK); training in data, AI system development, AI risks, and AI ethical principles targeting developers (TDEVEL); training in data for users (TUSER); and training in data, AI risks and AI ethical principles for auditors (TAUDIT). Focusing on figuring out whether the mentioned trainings are enablers of AI governance practices, the overall score of actions towards AI governance was considered as the outcome variable (actions at the strategic level, ancillary processes and practices, and practices that belong to the AI system development process) (GAIGOV(1)).

Delivering a consistency of 0.937173, the fuzzy QCA revealed that the combination characterized by training aimed at decision-makers, developers, and users showed 87.5 % (cases 8, 9, 13, 22, 24, 27, 28) of these cases with high scores for actions focused on AI governance. Such result indicates an association between training key stakeholders and higher stage implementation of AI governance, as argued by Calzada & Almirall, 2020; Micheli et al., 2020; Ruijer, 2021; Makarius et al., 2020; Herremans, 2021, Pinski et al., 2024; Schüller, 2022.

Table 7
Fuzzy QCA applied to Proposition 2.

Proposition	Combinations	Cases	Results	Coverage and Consistency
2	1. TDEVEL * ~ TAUDIT	27,19,20,28 14,25,26	GAIGOV (1) GAIGOV (0)	Cv: 0.452348 Cs: 0.927762
	2. TDEMAK * TDEVEL * TUSER	13,22,27,8,9,24,28	GAIGOV (1)	Cv: 0.494475 Cs: 0.937173

Table 8
Crisp-set Fuzzy QCA applied to Proposition 3.

Proposition	Combinations	Cases	Results
3	1. TUSEAI{1}*DTDEMAK{0}* DTDEVEL{0}	6,3	HAIGOV(1)
		1,4,2,10,15,18,21,23	HAIGOV(0)
	2. ACOD80{0}*TUSEAI{0}* DTDEMAK{1}	5,20,28	HAIGOV(1)
		22	HAIGOV(0)
	3. ACOD80{0}*TUSEAI{0}* DTDEVEL{1}	19,20,28	HAIGOV(1)
		22	HAIGOV(0)
	4. TUSEAI{1}*DTDEMAK{1}* DTDEVEL{1}	8,9,13,27,24,25	HAIGOV(1)
		26	HAIGOV(0)

The absence of auditor training revealed that focus on internal audits for AI systems has not been a priority for those organizations, although a small group has provided the four trainings.

4.3. Analysis of proposition 3

Proposition 3 was analyzed through a crisp-set QCA (Table 8) using the following dichotomous variables (Appendix A - Annex 6): training for decision-makers (DTDEMAK), training for developers (DTDEVEL), the organization's access to at least 80 % of their AI system code (ACOD80) (four of the five AI systems reported in the questionnaire), and more than three years of experience in AI system development (TUSEAI). For the outcome variable, a dichotomous variable was created to indicate the sum of the fuzzy scores of all practices for implementing AI governance (HAIGOV). HAIGOV = 1, if = GAIGOV≥60, and HAIGOV = 0 if GAIGOV<60.

Combination 1 — consisting of cases with more than three years in AI system production, without training for decision-makers and for AI system developers — was associated with low (not high) overall scores for all actions towards AI governance (cases 1, 2, 4, 10, 15, 18, 21, 23). At the other end, Combination 4 — comprising cases with more than three years of AI system production that have offered training to decision-makers and AI system developers — was associated with more advanced stages of the implementation of AI governance (cases 8, 9, 13, 24, 25, 27). Combinations 2 and 3 were not assertive enough to any conclusion.

It is worth observing that having (or not) access to the AI system codes, which implies outsourcing AI systems, did not impact Combinations 1 and 4. According to Combination 4, training decision-makers and developers can be associated with an advanced stage in implementing AI governance practices. And according to Combination 1, not training decision-makers or developers can be associated with lower stages in implementing AI governance. Both situations align with Ahn and Chen (2022) and Benfeldt et al. (2020). Preposition 3 reinforces the need to train key stakeholders even when the public organization outsources the development of AI systems.

5. Analysis of the interview responses and documents

The analysis of the interview responses and documents provided by the organizations was carried out while attempting to understand how practices and processes categorized in the research model were applied and used in the analysis of propositions.

5.1. Processes and practices for AI governance

In the context of actions, processes, and practices, many approaches were given and challenges were found in the path towards AI governance in the studied public organizations (Appendix A - Annex 7)(CNJ, 2020; LIAA-3R, 2022; Nagbøl & Müller, 2020; Nagbøl, Müller, & Krancher, 2021; Vero, 2019). Since all efforts towards AI governance are distributed at many levels of the organization hierarchy, it reflects the organization's culture and its risk appetite.

5.2. Government standards and guidelines

Along with the interviews, one observes that organizations whose governments have already established AI ethical principles guidelines for all their agencies have adopted those principles completely and, in a few cases, they have added details to their policies or strategies to customize the guidelines to their singularities (See "Policies and Guidelines for AI Ethical Principles" in Appendix A - Annex 7 and Appendix A - Annex 9). Similarly, in some cases, governments create agencies specialized in developing standards for processes and practices related to AI governance (See Appendix A - Annex 8). In both cases, those organizations saved time, money, and human resources, as one can observe in Appendix A - Annex 10.

Regarding the construction of standards and transfer of knowledge, it is worth highlighting a partnership between the public and private sectors established by the Finnish Government (Aurora, 2019) and an agreement between Nordic countries to implement best practices for AI systems with a focus on ethical issues (Nordic Council of Ministers, 2018).

5.3. Outsourcing

In 73.33 % of the organizations, third parties were hired or partnered with to develop at least part of their AI systems. Considering that public organizations are not self-sufficient to produce AI systems on the scale and with the level of expertise they require, according to Hickok (2022), Zick et al. (2024) and Coglianese (2024), outsourcing is also a driver to implement practices for AI governance (Appendix A - Annex 10). A few organizations were inspired by the World Economic Forum's model (WEF, 2020, 2020b) for outsourcing AI system development compliant with ethical principles.

5.4. Partnership between the business unit and the IT unit

Among the interviewees, there is the perception of the "business + IT" joint action as a strategy to minimize biases, provide transparency, and implement data governance. A second group of reports was provided by professionals who implemented the risk management process and the AI system development process, with artifacts filled out by the IT and business staff (Nagbøl et al., 2021). In Annex 10, one can find some of the interviewees' responses and comments regarding the "business + IT" partnership.

6. Merging the analyses

6.1. Associations found in the QCA

The proposition results are summarized as follows:

Proposition 1 results: An association was found between AI governance actions at the strategic level ("create an AI governance structure," "elaborate an AI strategy," "establish an AI policy," "implement an AI governance process," "establish an AI ethical principles code") and the existence of data-related processes (data governance, data quality management, personal data protection management), risk-related processes and practices (risk management process, stakeholder definitions, monitoring changes in the environment), security management processes, AI system development practices (rules representing ethical principles and ethical dilemmas, biases minimization, transparency, automatic monitoring, human oversight, and feedback collection). No association was found between the audit process and the strategic actions for AI governance.

Proposition 2 results: An association was found between advanced stages of AI governance implementation and training targeting decision-makers, AI system developers, and digital services users, confirming that training such stakeholders is a driver for implementing AI governance.

Proposition 3 results: In the context of organizations with more than three years of experience in AI system production, regardless of the make-or-buy decision, training decision-makers and developers is associated with an advanced stage of implementing all AI governance practices. And not training decision-makers or developers is associated with lower stages in implementing AI governance. Thus, regardless of the make-or-buy decision, training is a driver of AI governance implementation.

6.2. *Relevant contributions from the government to the whole public sector*

Interviews and documents made available by governments showed the benefits of having a central government body that produces clear and accessible guidelines with recommendations for AI system development, which confirms Mikalef et al.'s (2022b) and Schaefer et al.'s (2021) perceptions. Some guides made up the governments' portfolio of standards for promoting AI governance in agencies and departments under their responsibility (Australian Government, 2019a, 2019b; Norwegian Data Protection Authority, 2018; Government of United Kingdom, 2017, 2019b, 2020a, 2020b, 2020c, 2020d, 2020e, 2021a, 2021b, 2021c, 2021d, 2021e, 2022a, 2022b, 2022c, 2022d, 2022e; Information Commissioner's Office, 2020, 2021; Ekspertgruppen om dataetik, 2018; Balahur et al., 2022; German Federal Ministry for Economic Affairs and Energy, 2020; AI HLEG, 2019; Government of Canada, 2019a, 2019b, 2020a, 2020b, 2021a, 2021b, 2022a, 2022b; Leslie, 2019; National Institute of Advanced Industrial Science and Technology, 2022; European Union Agency for Cyber Security, 2021; WEF, 2020a, 2020b; C4IR Brasil, 2022; Switzerland Federal Council, 2021, Council of Europe, 2021, European Data Protection Board, 2022) (Appendix A - Annex 8). In addition to supplying knowledge that is lacking in many public organizations, these specialized institutions speed up implementation and promote a standard for concepts that facilitates communication among government departments. In a strategic context, some governments (German Federal Government, 2020;, Presidencia de

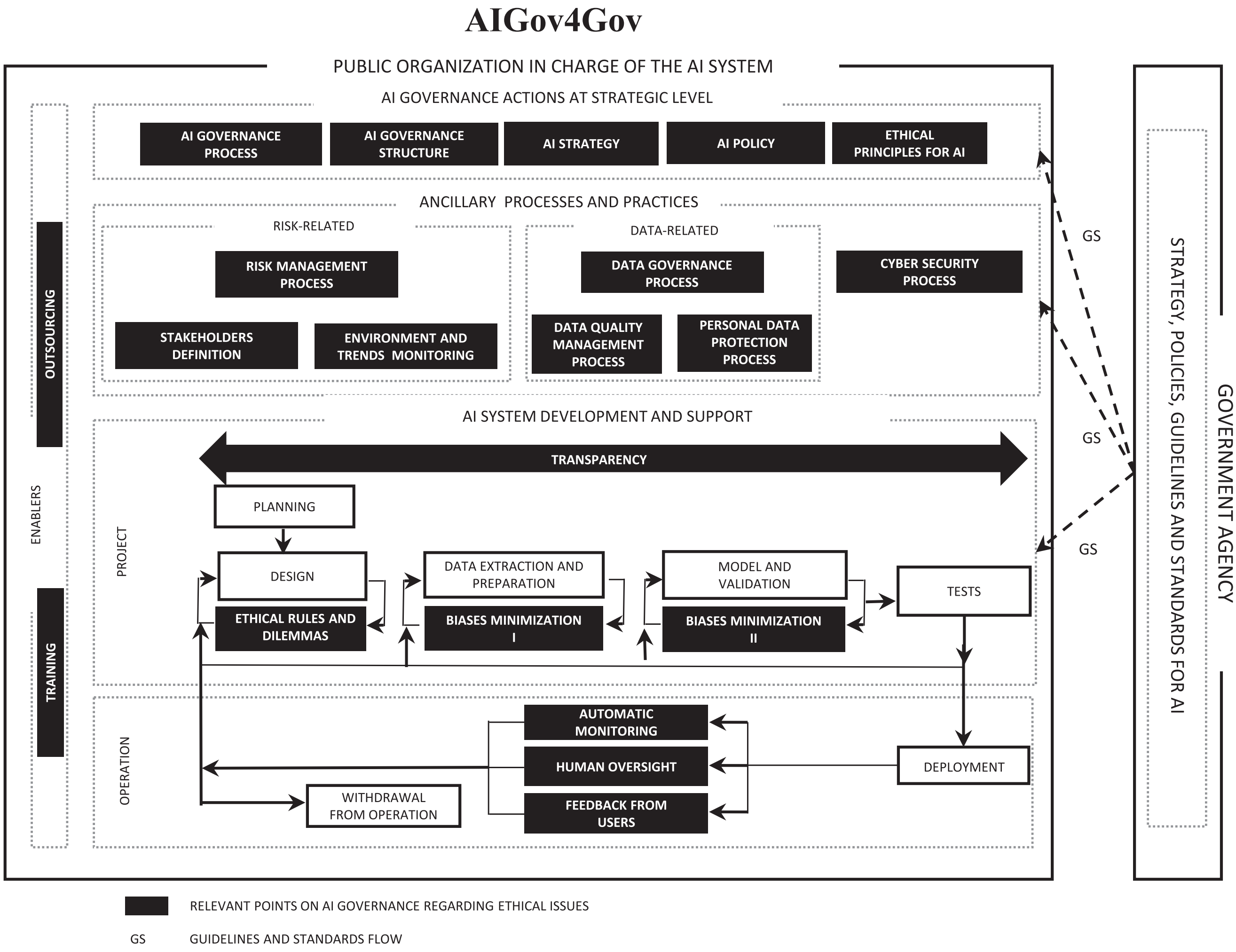

AIGov4Gov
PUBLIC ORGANIZATION IN CHARGE OF THE AI SYSTEM
AI GOVERNANCE ACTIONS AT STRATEGIC LEVEL
AI GOVERNANCE PROCESS
AI GOVERNANCE STRUCTURE
AI STRATEGY
AI POLICY
ETHICAL PRINCIPLES FOR AI
ANCILLARY PROCESSES AND PRACTICES
RISK-RELATED
RISK MANAGEMENT PROCESS
STAKEHOLDERS DEFINITION
ENVIRONMENT AND TRENDS MONITORING
DATA-RELATED
DATA GOVERNANCE PROCESS
DATA QUALITY MANAGEMENT PROCESS
PERSONAL DATA PROTECTION PROCESS
CYBER SECURITY PROCESS
AI SYSTEM DEVELOPMENT AND SUPPORT
TRANSPARENCY
PROJECT
PLANNING
DESIGN
ETHICAL RULES AND DILEMMAS
DATA EXTRACTION AND PREPARATION
BIASES MINIMIZATION I
MODEL AND VALIDATION
BIASES MINIMIZATION II
TESTS
OPERATION
AUTOMATIC MONITORING
HUMAN OVERSIGHT
FEEDBACK FROM USERS
DEPLOYMENT
WITHDRAWAL FROM OPERATION
ENABLERS
OUTSOURCING
TRAINING
GS
GS
GS
GOVERNMENT AGENCY
STRATEGY, POLICIES, GUIDELINES AND STANDARDS FOR AI
RELEVANT POINTS ON AI GOVERNANCE REGARDING ETHICAL ISSUES
GS GUIDELINES AND STANDARDS FLOW

Fig. 6. AIGov4Gov – AI governance framework proposed for public organizations. (Source: Self-elaboration)

la Nación, 2019; Australian Government, 2021; Ministério da Ciência Tecnologia e Inovação do Brasil, 2021; Government of Canada, 2022b; Danish Government, 2019; Government of the Republic of Estonia, 2019; Ministry of Economic Affairs and Employment of Finland, 2017, 2019; Ministero dello sviluppo economico, 2019; Japanese Strategic Council for AI Technology, 2017; Ekspertgruppen om dataetik, 2018; Government of the Grand Duchy of Luxembourg, 2018; Norwegian Ministry of Local Government and Modernisation, 2020; Government of United Kingdom, 2021e; Government of Sweden, 2020; European Commission, 2018a) assign AI strategic guidelines to an agency that develops and monitors the national AI strategy and establishes policy and ethical principles for AI systems (Appendix A - Annex 9). In both cases, public organizations can not only adopt the government's central strategic guidelines and standards but can also build their own versions in line with the government's general definitions.

6.3. *Drivers to AI governance implementation*

Combining the findings of Propositions 2 and 3 with the interviews, training key stakeholders and outsourcing were considered drivers of AI governance implementation in public organizations.

6.4. *AIGov4Gov framework*

Consolidating the QCA results and the interviews' results, a conceptual view was built in a multilevel approach encompassing practices and processes found in the sample aimed at the production of AI systems that considered ethical principles, drivers to implement AI governance, and the Government Central Agency contributions to the public sector: the AIGov4Gov framework (Fig. 6).

Reaching the strategic, tactical, and operational levels, AIGov4Gov is an AI governance model for public organizations. At the strategic level, there are actions aimed at elaborating an AI strategy, establishing an AI policy, establishing an AI ethical principles code, implementing an AI governance process, and creating an AI governance structure for said governance. To support those actions, ancillary processes and practices are combined for a) data governance, supported by data quality management and personal data protection management; b) AI-related risk mitigating using a risk management process supported by a stakeholders' definition and by an environment and social trends monitoring in line with an audit process; c) Cyber security management. The expected increase of legislation to regulate AI worldwide was considered when deciding to maintain the audit process within the framework. At the tactical and operational levels, AIGov4Gov provides practices for AI system development and maintenance processes with a focus on ethical principles and aligned to researchers' argument for an AI system development process based on agile methods, which use continuous loops in each phase (Laato et al., 2022), and which consider ethical principles in the loops (Leijnen et al., 2020; Lu et al., 2024). Right after planning, development takes place in several interactions during problem specification and with the representation of ethical principles and dilemmas. Then, practices for minimizing biases are applied in successive iterations during the data extraction and preparation phase, as well as during the model construction and validation stage, followed by testing. Soon after deployment, practices to follow-up AI systems are implemented in the complexity of the real environment through automatic monitoring, human oversight, and collection of user feedback.

As an enabler of AI governance implementation, AIGov4Gov provides outsourcing in addition to training for key stakeholders, like decision-makers, developers, and users, customized for their role in the implemented processes and practices. Auditor training can also be considered when aiming for a scenario where AI legislation is a comprehensive reality.

Also included in AIGov4Gov is the interaction between the public organization in charge of the AI system and the agency in its sphere of government (if any) in charge of AI strategy, AI policy, and AI ethical principles for AI systems applicable to organizations under its responsibility. In such a situation, the organization in charge of the AI system can adopt its centralized government strategic guidelines or adapt them while complying with them. Similarly, guidelines are provided with standards for risk management, data governance, data quality management, personal data protection management, cyber security management, minimization of biases, and transparency throughout the AI system development process (GS Flow in Fig. 6). When the organization decides to outsource, the AI system development process still requires the "business + IT" partnership to enable the ancillary processes and practices for AI governance.

7. Conclusions

This study investigated how public organizations have incorporated the guidelines presented by academia, international standards, and legislation for AI system development, considering ethical principles in their governance and management processes. The results confirmed the perception that AI governance requires a multilayer model with strategic-level actions that guide processes and ancillary AI governance practices in a combined action. All processes and practices designed in the research model were observed in the sample. However, data-related, risk-related, and AI system development processes and practices were prioritized in the sample for AI governance implementation. The cyber security management process received lower adherence, and audit processes were still seldom adopted at the time of data gathering when very few countries had approved laws to regulate AI. Regarding development in the project phase, organizations at a more advanced stage in AI governance have prioritized practices representing ethical principles and ethical dilemmas, transparency practices, or practices to minimize biases. When those AI systems are in operation, organizations at a more advanced stage in AI governance have implemented practices for automatic monitoring, human oversight, and collection of user feedback.

Training key stakeholders and outsourcing are enablers of an AI governance implementation. It was observed that organizations that have outsourced their AI system development have also trained managers as well as AI developers, and the lack of training of those professionals is associated with less advanced stages in AI governance.

One could observe the huge opportunity that government agencies have to promote AI governance by defining guidelines for all organizations under their responsibility or recommending standards for AI governance despite the long time required to deploy them. In the context of public organizations, it is worth mentioning the need for policymaking that combines an internal multilayered approach with a continuous alignment with the government guidelines and standards. The findings also corroborated to a proposed framework — AIGov4Gov — encompassing combined processes and practices to establish AI governance in a public organization.

8. Limitations and agenda

Limitations found in this research:

a) Despite the broad and systematic process of obtaining the sample, this research carried out analyses in countries and organizations that published their AI systems, made contacts available, and agreed to participate in the research. Therefore, despite efforts to include representatives from all countries with a high level of AI system production, the sample does not follow the global AI ranking proportions, nor does it have a balanced representation of each continent.

b) The research focused on capturing the existence of practices and processes but did not delve into each process and its maturity model. Thus, each participant had his own perception regarding whether a practice/process was being implemented completely or partially. For the same reason, the research did not make a deeper analysis of the quality of the training offered to users, developers, decision-makers, and auditors.

c) Similar to the previous item, the government AI standards found in the sample were not classified considering their maturity.

d) The research encompassed only practices and processes representing all efforts to implement AI governance. Impacts on corporate governance and e-government were out of the scope.

e) The AI systems presented by the organizations in the sample did not consider generative AI, probably because the gathering criteria were systems that were in operation and were already part of the organization's official portfolio when the data were collected.

As an exploratory and descriptive study, this research paves the way for an agenda of new investigations that go deeper into the findings regarding each proposition, as well as an investigation of how practices and processes studied under the AI governance effort would impact corporate governance and e-government. Finally, a space is opened for deepening the AI system lifecycle through a maturity model for AI system development and support focusing on ethical principles.

9. Contribution

This research presents itself as innovative in terms of content, as it addressed the gap highlighted by Mäntymäki et al. (2022) and Mikalef, Conboy, et al. (2022) in the empirical knowledge of how organizations have interpreted and incorporated AI system development best practices into their processes and practices. The research has innovated by using crisp-set QCA, fuzzy QCA, and content analyses as it addressed the gaps presented by Zuiderwijk et al. (2021). Therefore, the following contributions to managers and researchers can be summarized: a) identification of how processes and practices aimed at applying ethical principles in AI system development have been combined and internalized in the governance and management models of public organizations; b) identification of how AI governance enablers have been used by public organizations; c) how a central government agency can booster AI governance in government agencies; d) a framework for AI governance in public organizations, in which processes and practices are articulated at the strategic, tactical, and operational levels in AI system production that consider ethical principles.

CRediT authorship contribution statement

Patricia Gomes Rêgo de Almeida: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Carlos Denner dos Santos Júnior:** Supervision, Writing- original draft, Investigation, Visualization, Validation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.giq.2024.102003.

References

Aasi, P., Rusu, L., & Han, S. (2014). The influence of culture on IT governance: A literature review. In *47th Hawaii international conference on system sciences* (pp. 4436–4445). https://doi.org/10.1109/HICSS.2014.546

Abdollahi, B., & Nasraoui, O. (2018). Transparency in Fair machine learning: The case of explainable recommender systems. In J. Zhou, & F. Chen (Eds.), *Human and Machine Learning. Human–Computer Interaction Series*. Cham: Springer. https://doi.org/10.1007/978-3-319-90403-0_2.

Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management, 49*(July), 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

Aburachid, L. M. C., & Greco, P. J. (2011). Validação de conteúdo de cenas do teste de conhecimento tático no tênis. *Estudos de Psicologia. Campinas, 28*(2), 261–267.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access Review, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Agarwal, L.. Defining organizational AI governance and ethics. Available at SSRN: https://ssrn.com/abstract=4553185.

Ahn, M. J., & Chen, Y. (2022). Digital transformation toward AI-augmented public administration: The perception of government employees and the willingness to use AI in government. *Government Information Quarterly, 39), Issue 2*. https://doi.org/10.1016/j.giq.2021.101664

AI HLEG. (2019). Ethics guidelines for trustworthy AI. European Commission. Retrieved from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Aiken, C. (2021). *Classifying AI systems. Center for Security and Emerging Technology*. Georgetown University. Retrieved from https://cset.georgetown.edu/publication/classifying-ai-systems/ Accessed September 23, 2024.

Alhosani, K., & Alhashmi, S. M. (2024). Opportunities, challenges, and benefits of AI innovation in government services: A review. *Discover Artificial Intelligence, 4*, 18. https://doi.org/10.1007/s44163-024-00111-w

Alshahrani, A., Dennehy, D., & Mäntymäki, M. (2021). An attention-based view of AI assimilation in public sector organizations: The case of Saudi Arabia. *Government Information Quarterly, 39), Issue 4*. https://doi.org/10.1016/j.giq.2021.101617

Anderson, M., & Anderson, S. L. (2018). Geneth: A general ethical dilemma analyzer. De Gruiter. Paladyn. *J. Behav. Robot., 9*, 337–357. https://doi.org/10.1515/pjbr-2018-0024

Andrews, L. (2018). Public administration, public leadership and the construction of public value in the age of the algorithm and 'big data'. *Public Administration, 97*(2), 296–310. https://doi.org/10.1111/padm.12534

Arrieta, A. B., Díaz-Rodríguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for artificial intelligence and digital technologies. *International Journal of Information Management, 62*. https://doi.org/10.1016/j.ijinfomgt.2021.102433

Ashokan, A., & Haas, C. (2021). Fairness metrics and bias mitigation strategies for rating predictions. *Information Processing & Management, 58*(5), Article 102646. ISSN 0306–4573 https://doi.org/10.1016/j.ipm.2021.102646.

Aurora, A. I. (2019). Aurora AI - Towards a human Centric Society. Retrieved from https://vm.fi/documents/10623/1464506/AuroraAI+development+and+implementation+plan+2019%E2%80%932023.pdf.

Australian Government. (2019). *Artificial intelligence – Australia's ethics framework*. Innovation and Science: Department of Industry. Retrieved from https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework Accessed September 23, 2024.

Australian Government. (2019b). Australian Privacy Principles Guidelines. Office of Australian Information Commissioner. Retrieved from https://www.oaic.gov.au/_data/assets/pdf_file/0009/1125/app-guidelines-july-2019.pdf Accessed September 23, 2024.

Australian Government. (2021). Australian's Artificial Intelligence Action Plan. Retrieved from https://www.industry.gov.au/publications/australias-artificial-intelligence-action-plan.

Awad, E., Dsouza, S., Bonnefon, J. F., Shariff, A., & Rahwan, I. (2020). Crowdsourcing moral machines. *Communications of the ACM, 63*(3), 48–55. https://doi.org/10.1145/3339904

Baeza-Yates, R. (2018). Bias on the web. *Communications of the ACM, 61*(6), 54–61. https://doi.org/10.1145/3209581

Balahur, A., Jenet, A., Torres, I., Charisi, V., Ganesh, A., Griesinger, C. B., Maurer, P., Mian, L., Salvi, M., Scalzo, S., Sol er Garrido, J., Taucer, F., & Tolan, S. (2022). *Data quality requirements for inclusive, non-biased and trustworthy AI. Putting-Science-Into-Standards*. Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/365479

Bench-Capon, T., & Modgil, S. (2017). Norms and value based reasoning: Justifying compliance and violation. *Artificial Intelligence & Law Review, 25*, 29–64. https://doi.org/10.1007/s10506-017-9194-9

Benfeldt, O., Persson, J. S., & Madsen, S. (2020). Data governance as a collective action problem. *Information Systems Frontiers, 22*, 299–313. https://doi.org/10.1007/s10796-019-09923-z

Betarelli-Júnior, A. A., & Ferreira, S. F. (2018). *Introdução à Análise Qualitativa Comparativa e aos Conjuntos Fuzzi (FSQCA)*. Enap: Brasilia.

Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: Some technical and ethical challenges. *Ethics and Information Technology, 20*, 41. https://doi.org/10.1007/s10676-018-9444-x

Bonsón, E., Lavorato, D., Lamboglia, R., & Mancini, D. (2021). Artificial intelligence activities and ethical approaches in leading listed companies in the European Union. *International Journal of Accounting Information Systems, 43*. https://doi.org/10.1016/j.accinf.2021.100535, 1467–0895.

Booth, J., Metz, D. W., Tarkhanyan, D. A., & Cheruvu, S. (2023). Machine learning security and trustworthiness. In *Demystifying intelligent multimode security systems*. Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-8297-7_5.

Boyd, R., & Holton, R. J. (2018). Technology, innovation, employment and power: Does robotics and artificial intelligence really mean social transformation? *Journal of Sociology, 54*(3), 331–345.

Breier, J., Baldwain, A., Balinsky, & Liu, Y. (2020). Risk Management for Machine Learning Security. *arXiv*. https://doi.org/10.48550/arXiv.2012.04884, 2012.04884v1 [cs.CR].

Buiten, C. M. (2019). Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation, 10*(1), 41–59. https://doi.org/10.1017/err.2019.8

Butterworth, M. (2018). The ICO and artificial intelligence: The role of fairness in the GDPR framework. *Computer Law and Security Review, 34*, 257–268. https://doi.org/10.1016/j.clsr.2018.01.004

C4IR Brasil. (2022). Guia de Contratações Públicas de Inteligência Artificial. Centro para a 4ª Revolução Industrial. Retrieved from https://c4ir.org.br/wp-content/uploads/2022/11/1648128585465GUIA-DE-CONTRATACOES-PUBLICAS-DE-AI_C4IR_v4.pdf Accessed September 23, 2024.

Calzada, I., & Almirall, E. (2020). Data ecosystems for protecting European citizens' digital rights. *Transforming Government: People, Process and Policy, 14*(2), 133–147. https://doi.org/10.1108/TG-03-2020-0047

Carretero, A., Gualo, F., Caballero, I., & Piattini, M. (2017). *MAMD 2.0: Environment for data quality processes implantation based on ISO 8000-6X and ISO/IEC 33000. 54* pp. 139–151). Elsevier BV. https://doi.org/10.1016/j.csi.2016.11.008

Carter, D. (2020). (2020). Regulation and ethics in artificial intelligence and machine learning technologies: Where are we now? Who is responsible? Can the information professional play a role? *Business Information Review, 37*(2), 60–68. https://doi.org/10.1177/0266382120923962

Cerka, P., Grigiene, J., & Sirbikyte, G. (2017). Is it possible to grant legal personality to artificial intelligence software systems? *Computer Law and Security Review, 33*(5), 685–699.

Chen, T., Liu, J., Xiang, Y., Niu, W., Tong, E., & Han, Z. (2019). Adversarial attack and defense in reinforcement learning-from AI security view. *Cybersecurity, 2*, 1–22. https://doi.org/10.1186/s42400-019-0027-x

Chen, X., & Deng, Y. (2022). An evidential software risk evaluation model. *Mathematics, 10*, 2325. https://doi.org/10.3390/math10132325

Cihon, P., Maas, M. M., & Kempo, L. (2020). Should artificial intelligence governance be centralized?: Design lessons from history. In *AAAI/ACM conference on AI, ethics, and society (AIES '20)* (pp. 228–234). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3375627.3375857.

CNJ. (2020). Resolução N° 332 de 21/08/2020. Retrieved from https://atos.cnj.jus.br/atos/detalhar/3429 Accessed September 23, 2024.

Codá, R. C., Farias, J. S., & Dias, C. (2022). Interactive value formation and lessons learned from Covid-19: The Brazilian case. *Journal of Quality Assurance in Hospitality and Tourism*. https://doi.org/10.1080/1528008X.2022.2135057

Coglianese, C. (2024). Procurement and artificial intelligence. In *Handbook on public policy and artificial intelligence* (pp. 235–248). Edward Elgar Publishing. https://doi.org/10.4337/9781803922171.00026.

Council of Europe. (2021). Guidelines on facial recognition. Consultative Committee of the Convention for the protection of individuals with regard to automatic processing of personal data. Convention 108. Retrieved from https://rm.coe.int/guidelines-facial-recognition-web-a5-2750-3427-6868-1/1680a31751 Accessed September 23, 2024.

Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N., Abate, A., Halpern, J., Barrett, C., Zhao, D., Zhi Xuan, T., Wing, J., & Tenenbaum, J. (2024). *Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems*. arXiv:2405.06624v3 [cs.AI]. https://doi.org/10.48550/arXiv.2405.06624.

Danish Government. (2019). *National Strategy for Artificial Intelligence*. Ministry of Finance and Ministry of Industry, Business and Financial Affairs. Retrieved from https://en.digst.dk/strategy/the-danish-national-strategy-for-artificial-intelligence/ Accessed September 23, 2024.

Danish Government. (2020). Lov om ændring af årsregnskabsloven. (Krav om rapportering af dataetik). Retrieved from https://www.retsinformation.dk/eli/lta/2020/741 Accessed September 23, 2024.

Das, A. (2020). *Opportunities and challenges in explainable artificial intelligence 9XAI: A survey*. arXiv:2006.11371 [cs.CV]. https://doi.org/10.48550/arXiv.2006.11371.

Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., & Cruz, F. (2021). Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence, 299, 103525*. https://doi.org/10.1016/j.artint.2021.103525. ISSN 0004-3702.

De Almeida, P. G. R., dos Santos, C. D., & Farias, J. S. (2021). Artificial intelligence regulation: A framework for governance. *Ethics and Information Technology, 23*, 505–525. https://doi.org/10.1007/s10676-021-09593-z

De Oliveira, T. F. (2019). *Avaliação das Práticas de Auditoria Interna da Secretaria Federal de Controle Interno da CGU sob a Ótica da Auditoria Baseada em Riscos*. Brasil: Controladoria Geral da União. Retrieved from https://revista.cgu.gov.br/Revista_da_CGU/article/view/73/pdf_60 Accessed September 23, 2024.

De Silva, D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. *Patterns, 3*(6), Article 100489. https://doi.org/10.1016/j.patter.2022.100489

Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems, 77*, 1–14. ISSN 0921-8890, https://doi.org/10.1016/j.robot.2015.11.012.

Dias, O. C. (2011). *Análise Qualitativa Comparativa (QCA) Usando Conjuntos Fuzzy – Uma Abordagem Inovadora Para Estudos Organizacionais no Brasil*. Rio de Janeiro: XXXV Encontro da ANPAD.

Dignum, V. (2019). AI is multidisciplinar. *AI Matters, 5*(4), 19–21. https://doi.org/10.1145/3375637.3375644

Djeffal, C. (2018). *Sustainable AI Development (SAID): On the Road to More Access to Justice*. https://doi.org/10.2139/ssrn.3298980

Duijm, N. J. (2015). Recommendations on the use and design of risk matrices. *Safety Science, 76*, 21–31. ISSN 0925-7535. https://doi.org/10.1016/j.ssci.2015.02.014

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., … Williams, M. D. (2021). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57*, Article 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Ee, S., O'Brien, J., Williams, Z., El-Dakhakhni, A., Aird, M., & Lintz, A. (2024). Adapting cybersecurity frameworks to manage frontier AI risks: A defense-in-depth approach. *arXiv*. https://doi.org/10.48550/arXiv.2408.07933, 2408.07933v1 [cs.CY].

Eggers, S., & Sample, C. (2020). *Vulnerabilities in artificial intelligence and machine learning applications and data*. Idaho National Laboratory. US. Retrieved from https://inldigitallibrary.inl.gov/sites/sti/sti/Sort_57369.pdf Accessed September 23, 2024.

Eke, D. O., Chintu, S. S., & Wakunuma, K. (2023). Towards shaping the future of responsible AI in Africa. In D. O. Eke, K. Wakunuma, & S. Akintoye (Eds.), *Responsible AI in Africa. Social and cultural studies of robots and AI*. Cham: Palgrave Macmillan. https://doi.org/10.1007/978-3-031-08215-3_8.

Ekspertgruppen om dataetik. (2018). Data i menneskets tjeneste Anbefalinger fra Ekspertgruppen om dataetik. Retrieved from https://www.em.dk/media/12013/ekspertgruppens-afrapportering-inkl-anbefalinger_final-a.pdf Accessed September 23, 2024.

Erlina, E., Nasution, A. A., Yahy, I., & Atmanegara, A. W. (2020). The role of risk based internal audit in improving audit quality. Erlina, Abdillah Arif Nasution, Idhar Yahya and Agung Wahyudhi Atmanegara, The Role of Risk Based Internal Audit in Improving Audit Quality. *International Journal of Management, 11*(12), 299–310.

European Commission. (2018). AI Watch. Retrieved from https://ai-watch.ec.europa.eu/index_en Accessed September 23, 2024.

European Commission. (2018a). Artificial Intelligence for Europe. Retrieved from https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence Accessed September 23, 2024.

European Commission. (2021a). Regulation of the European Parliament and of the Council Laying down harmonized rules on artificial intelligence. Artificial Intelligence Act and amending certain union legislative acts. Brussels. Retrieved from https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF Accessed September 23, 2024.

European Commission. (2021b). Selected AI cases in the public sector. Joint Research Centre [Dataset] PID. Retrieved from http://data.europa.eu/89h/7342ea15-fd4f-4184-9603-98bd87d8239a Accessed September 23, 2024.

European Data Protection Board. (2022). Guidelines 05/2022 on the use of facial recognition technology in the area of law enforcement. Version 1.0 Retrieved from https://edpb.europa.eu/our-work-tools/documents/public-consultations/2022/guidelines-052022-use-facial-recognition_en Accessed September 23, 2024.

European Parliament. (2022a). *Draft Report on the proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending Certain Union Legislative Acts (COM2021/0206 – C9–0146/2021–2021/0106(COD))*.

European Parliament. (2022b). Regulatory divergences in the draft AI act: Differences in public and private sector obligations, Study, European Parliamentary Research Service (EPRS), Brussels. Retrieved from https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729507/EPRS_STU(2022)729507_EN.pdf Accessed September 23, 2024.

European Union Agency for Cyber Security. (2021). *Securing Machine Learning Algorithms*. ISBN: 978–92–9204-543-2. https://doi.org/10.2824/874249.

European Union Agency for Cyber Security. (2022). Risk Management Standards – Analysis of standardisaton requeriments in supporting of cybersecurity policy. Retrieved from https://www.enisa.europa.eu/publications/risk-management-standards Accessed September 23, 2024.

FCT. (2021). Research in Data Science and Artificial Intelligence applied to Public Administration. Retrieved from https://www.fct.pt/wp-content/uploads/2022/06/Brochura_ResearchinDataScienceandAIappliedtoPA.pdf Accessed September 23, 2024.

Fernandes, G., Domingues, J., Tereso, A., & Pinto, E. (2021). A Stakeholders' perspective on risk Management for Collaborative. University-industry R&D programs. *Procedia Computer Science, 181*, 110–118. ISSN 1877-0509. https://doi.org/10.1016/j.procs.2021.01.110

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI* (p. 1). Berkman Klein Center Research Publication.

Freitas, V. S., & Neto, F. B. (2016). Qualitative Comparative Analysis (QCA): usos e aplicações do método. *Revista Política Hoje, 2ª Edição, 24*, 103–117.

Gerkensmeier, B., & Ratter, B. M. W. (2018). Governing coastal risks as a social process—Facilitating integrative risk management by enhanced multi-stakeholder collaboration. *Environmental Science & Policy, 80*, 144–151. ISSN 1462-9011. https://doi.org/10.1016/j.envsci.2017.11.011

German Federal Government. (2020). Artificial intelligence strategy of the German Federal Government. Retrieved from https://www.ki-strategie-deutschland.de/home.html Accessed September 23, 2024.

German Federal Ministry for Economic Affairs and Energy. (2020). German Standardization Roadmap on Artificial Intelligence. Retrieved from https://www.din.de/resource/blob/772610/e96c34dd6b12900ea75b460538805349/normungsroadmap-en-data.pdf Accessed September 23, 2024.

Gianni, R., Lehtinen, S., & Nieminen, M. (2022). Governance of responsible AI: From ethical guidelines to cooperative policies. *Frontiers in Computer Science, 4*, Article 873437. https://doi.org/10.3389/fcomp.2022.873437

González, F., Ortiz, T., & Ávalos, R. S. (2020). Responsible use of AI for public policy: Data science toolkit. In *OECD e IDB*. Retrieved from https://publications.iadb.or g/publications/english/document/Responsible-use-of-AI-for-public-policy-Data-science-toolkit.pdf Accessed September 23, 2024.

Government of Canada. (2019a). Directive on Identity Management. Retrieved from https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=16577 Accessed September 23, 2024.

Government of Canada. (2019b). Directive on Security Management. Retrieved from https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32611 Accessed September 23, 2024.

Government of Canada. (2020). Guidelines on Service and Digital. Retrieved from https://www.canada.ca/en/government/system/digital-government/guideline-service-digital.html#ToC4_5 Accessed September 23, 2024.

Government of Canada. (2020a). Algorithmic Impact Assessment. Retrieved from https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html Accessed September 23, 2024.

Government of Canada. (2021). Directive on Automated Decision-making. Retrieved from https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592 Accessed September 23, 2024.

Government of Canada. (2021d). Levels of security. National Security and defense. Retrieved from https://www.tpsgc-pwgsc.gc.ca/esc-src/protection-safeguarding/niveaux-levels-eng.html.

Government of Canada. (2022a). Responsible use of artificial intelligence – Our guiding principles. Retrieved from https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html#toc1 Accessed September 23, 2024.

Government of Canada. (2022). Pan-Canadian Artificial Intelligence Strategy. Retrieved from https://ised-isde.canada.ca/site/ai-strategy/en Accessed September 23, 2024.

Government of India. (2020). Artificial Intelligence – Use Case Compendium. Ministry of Housing and Urban Affairs. Retrieved from https://iccc.smartcities.gov.in/pdf/AI-Use-Case-Compendium.pdf Accessed September 23, 2024.

Government of Sweden. (2020). AI Sweden. Retrieved from https://www.ai.se/en/news/swedish-government-establishes-new-ai-commission Accessed September 23, 2024.

Government of the Grand Duchy of Luxembourg. (2018). Artificial Intelligence: a strategic view for Luxembourg. Retrieved from https://innovative-initiatives.public.lu/initiatives/artificial-intelligence-strategic-vision-luxembourg Accessed September 23, 2024.

Government of the Republic of Estonia. (2019). Kratt – Estonian Artificial Intelligence Deployment. Retrieved from https://f98cc689-5814-47ec-86b3-db505a7c3978.filesusr.com/ugd/7df26f_27a618cb80a648c38be427194affa2f3.pdf Accessed September 23, 2024.

Government of United Kingdom. (2017). Public sector use of the cloud. Retrieved from https://www.gov.uk/guidance/public-sector-use-of-the-public-cloud Accessed September 23, 2024.

Government of United Kingdom. (2019). Understanding artificial intelligence ethics and safety. Retrieved from https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety Accessed November 4, 2022.

Government of United Kingdom. (2020a). Data Ethics framework. Government Digital Service. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/923108/Data_Ethics_Framework_2020.pdf Accessed September 23, 2024.

Government of United Kingdom. (2020b). Guidelines for AI procurement. Retrieved from https://www.gov.uk/government/publications/guidelines-for-ai-procurement/guidelines-for-ai-procurement Accessed September 23, 2024.

Government of United Kingdom. (2020c). *Guidelines for AI procurement. A summary of best practices addressing specific challenges of acquiring Artificial Intelligence in the public sector.* UK Office for Artificial Intelligence. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/990469/Guidelines_for_AI_procurement.pdf Accessed September 23, 2024.

Government of United Kingdom. (2020d). The Government Data Quality Framework. Retrieved from https://www.gov.uk/government/publications/the-government-data-quality-framework/the-government-data-quality-framework-guidance Accessed February 26, 2023.

Government of United Kingdom. (2020e). A guide to using artificial intelligence in the public sector. https://www.gov.uk/government/publications/a-guide-to-using-artificial-intelligence-in-the-public-sector Accessed September 23, 2024.

Government of United Kingdom. (2021a). Ethics, Transparency and Accountability Framework for Automated Decision-Making. https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making/ethics-transparency-and-accountability-framework-for-automated-decision-making Accessed September 23, 2024.

Government of United Kingdom. (2021b). Algorithmic transparency template. Retrieved from https://www.gov.uk/government/publications/algorithmic-transparency-template/algorithmic-transparency-template Accessed September 23, 2024.

Government of United Kingdom. (2021c). Algorithmic Transparency Standard. https://www.gov.uk/government/collections/algorithmic-transparency-standard Accessed December 4, 2022.

Government of United Kingdom. (2021d). Using personal data in your business or other organization. Retrieved from https://www.gov.uk/guidance/using-personal-data-in-your-business-or-other-organisation#data-protection-and-gdpr Accessed September 23, 2024.

Government of United Kingdom. (2021e). Data Protection Impact Assessments. Retrieved from https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments Accessed September 23, 2024.

Government of United Kingdom. (2022a). *Standard. Food Standards Agency*. Food Hygiene Rating Scheme – AI. Retrieved from https://www.gov.uk/government/publications/food-standards-agency-food-hygiene-rating-scheme-ai/food-standards-agency-food-hygiene-rating-scheme-ai Accessed September 23, 2024.

Government of United Kingdom. (2022b). Ethics self-assessment tool. UK statistics authority. Retrieved from https://uksa.statisticsauthority.gov.uk/the-authority-board/committees/national-statisticians-advisory-committees-and-panels/national-statisticians-data-ethics-advisory-committee/ethics-self-assessment-tool/.

Government of United Kingdom. (2022c). Equality Impact Assessment. Retrieved from https://www.gov.uk/government/consultations/emergency-evacuation-information-sharing/equality-impact-assessment Accessed September 23, 2024.

Government of United Kingdom. (2022d). Service Standard. Retrieved from https://www.gov.uk/service-manual/service-standard Accessed September 23, 2024.

Government of United Kingdom. (2022e). Data Sharing Governance Framework. Retrieved from https://www.gov.uk/government/publications/data-sharing-governance-framework/data-sharing-governance-framework Accessed February 26, 2023.

Gräf, M., Mehler, M., & Ellenrieder, S. (2024). AI strategy in action: A case study on make-or-buy for AI-based services. In *Publications of Darmstadt Technical University, Institute for Business Studies (BWL) 146709, Darmstadt Technical University, Department of Business Administration, Economics and Law*. Institute for Business Studies (BWL).

Gu, T., Dolan-Gavitt, B., & Garg, S. (2019). BadNets: Identifying vulnerabilities in machine learning model supply. *arXiv*. https://doi.org/10.48550/arXiv.1708.06733, 1708.06733v2 [cs.CR].

Gutierrez, C. I., & Marchant, G. (2021). *A global perspective of soft law programs for the governance of artificial intelligence*. Sandra Day O`Connor College of Law. Arizona State University.

Haneem, F., Kama, N., Taskin, N., Pauleen, D., & Abu Bakar, N. A. (2019). Determinants of master data management adoption by local government organizations: An empirical study. *International Journal of Information Management, 45*, 25–43. https://doi.org/10.1016/j.ijinfomgt.2018.10.007

Hernández-Nieto, R. (2002). *Contributions to statistical analysis*. Mérida: Universidad de Los Andes.

Herremans, D. (2021). aiSTROM–A roadmap for developing a successful AI strategy. IEEE. *Access, 9*, 155826–155838. https://doi.org/10.1109/ACCESS.2021.3127548

Hickman, E., & Petrin, M. (2021). Trustworthy AI and corporate governance: The EU's ethics guidelines for trustworthy artificial intelligence from a company law perspective. *European Business Organization Law Review, 22*, 593–625. https://doi.org/10.1007/s40804-021-00224-0

Hickok, M. (2022). Public procurement of artificial intelligence systems: New risks and future proofing. *AI & SOCIETY, 1-15*. https://doi.org/10.1007/s00146-022-01572-2

Holmström, J. (2022). From AI to digital transformation: The AI readiness framework. *Business Horizons, 65*(3), 329–339. https://doi.org/10.1016/j.bushor.2021.03.006

Hopster, J. (2021). What are socially disruptive technologies? *Technology in Society, 67*, Article 101750. https://doi.org/10.1016/j.techsoc.2021.101750

IEEE. (2019). Ethically aligned design. In *Committees of the IEEE global initiative on ethics of autonomous and intelligent systems. 2nd version*. Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf Accessed September 23, 2024.

IEEE. (2020). P7010 - Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems. Retrieved from https://standards.ieee.org/ieee/7010/7718/ Accessed September 23, 2024.

IEEE. (2021a). P7000 - Model Process for Addressing Ethical Concerns During System Design. Retrieved from https://standards.ieee.org/ieee/7000/6781/ Accessed September 23, 2024.

IEEE. (2021b). P7001 - Transparency of Autonomous Systems. Retrieved from https://standards.ieee.org/ieee/7001/6929/ Accessed September 23, 2024.

IEEE. (2021c). P7005 - Standard for Transparent Employer Data Governance Accessed September 23, 2024 https://standards.ieee.org/ieee/7005/7014/ Accessed September 23, 2024.

IEEE. (2021d). P7007 - Ontological Standard for Ethically Driven Robotics and Automation Systems. Retrieved from https://standards.ieee.org/ieee/7007/7070/ Accessed September 23, 2024.

IEEE. (2022). P7002 - Data Privacy Process. Retrieved from https://standards.ieee.org/ieee/7002/6898 Accessed September 23, 2024.

Information Commissioner’’s Office. (2020). Big Data, artificial intelligence, machine learning and data protection. Version 2.2. Retrieved from https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf Accessed September 23, 2024.

Information Commissioner’’s Office. (2021). International data transfers. Retrieved from https://ico.org.uk/for-organisations/dp-at-the-end-of-the-transition-period/data-protection-and-the-eu-in-detail/the-uk-gdpr/international-data-transfers/ Accessed September 23, 2024.

IPS-X. (2021). IPS-X survey. European cases. Retrieved from https://ipsoeu.github.io/ips-explorer/case/.

ISO. (2021a). ISO/IEC 24027 - Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making. Retrieved from https://www.iso.org/standard/77607.html Accessed September 23, 2024.

ISO. (2021b). ISO/IEC 24372 - Information technology — Artificial intelligence (AI) — Overview of computational approaches for AI systems. Retrieved from https://www.iso.org/standard/78508.html Accessed September 23, 2024.

ISO. (2021c). ISO/IEC 24668 - Information technology — Artificial intelligence — Process management framework for big data analytics. Retrieved from https://www.iso.org/standard/78368.html Accessed September 23, 2024.

ISO. (2022a). ISO/IEC 38507 - Information Technology – Governance implications of the use of artificial intelligence by organizations. Retrieved from https://www.iso.org/standard/56641.html Accessed September 23, 2024.

ISO. (2022b). ISO/IEC 23894 – Information Technology – Risk management. Retrieved from https://www.iso.org/standard/77304.html Accessed September 23, 2024.

IT Governance Institute. (2003). Board Briefing on IT Governance. 2nd ed. Retrieved from http://www.gti4u.es/curso/material/complementario/itgi_2003.pdf Accessed September 23, 2024.

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly, 37*(3), Article 101493. https://doi.org/10.1016/j.giq.2020.101493

Japanese Strategic Council for AI Technology. (2017). Artificial Intelligence Technology Strategy. Retrieved from https://ai-japan.s3-ap-northeast-1.amazonaws.com/7116/0377/5269/Artificial_Intelligence_Technology_StrategyMarch2017.pdf Accessed September 23, 2024.

Jing, H., Wei, W., Zhou, C., & He, X. (2021). An Artificial Intelligence Security Framework. In *Journal of Physics: Conference Series, Volume 1948, The 2021 2nd International Conference on Internet of Things. Artificial Intelligence and Mechanical Automation (IoTAIMA 2021), 14–16, Hangzhou, China*.

Kale, A., Nguyen, T., Harris, F. C., Li, C., Zhang, J., & Ma, X. (2022). Provenance documentation to enable explainable and trustworthy AI: A literature review. *Data Intelligence, 1-41*. https://doi.org/10.1162/dint_a_00119

Kazim, E., & Koshiyama, A. S. (2021). A high-level overview of AI ethics, patterns, 2(9). *ISSN, 100314*, 2666–3899. https://doi.org/10.1016/j.patter.2021.100314

Khatri, V. (2016). Managerial work in the realm of the digital universe: The role of the data triad. *Business Horizons, 59*(6), 673–688. https://doi.org/10.1016/j.bushor.2016.06.001

Kitsios, F., & Kamariotou, M. (2021). Artificial intelligence and business strategy towards digital transformation: A research agenda. *Sustainability, 13*. https://doi.org/10.3390/su13042025

Korjani, M. M., & Mendel, J. M. (2012). Fuzzy set qualitative comparative analysis (fsQCA): Challenges and applications. In *Annual Meeting of the North American Fuzzy Information Processing Society (NAFIPS), 1-6*. https://doi.org/10.1109/NAFIPS.2012.6291026

Krippendorff, K. (2013). *Content analysis – An introduction to its methodology* (3td ed.). Sage.

Kuziemski, M., & 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), Article 101976. https://doi.org/10.1016/j.telpol.2020.101976

Laato, S., Birkstedt, T., Mäantymäki, M., Minkkinen, M., & Mikkonen, T. (2022). AI governance in the system development life cycle: Insights on responsible machine learning engineering. In *In proceedings of the 1st international conference on AI engineering: Software engineering for AI* (pp. 113–123). https://doi.org/10.1145/3522664.3528598

Labadie, C., Legner, C., Eurich, M., & Fadler, M. (2020). FAIR enough? Enhancing the usage of Enterprise data with data catalogs. IEEE. In *22nd conference on business informatics (CBI)* (pp. 201–210). https://doi.org/10.1109/CBI49978.2020.00029

Leavy, S., O'Sullivan, B., & Siapera, E. (2020). Data, power and Bias in artificial intelligence. *ArXiv*. https://doi.org/10.48550/arXiv.2008.07341. abs/2008.07341.

Leijnen, S., Aldewereld, H., van Belkom, R., Bijvank, R., & Ossewaarde, R. (2020). *An agile framework for trustworthy AI. In NeHuAI@ ECAI* (pp. 75–78).

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. *The Alan Turing Institute.*. https://doi.org/10.5281/zenodo.3240529

LIAA-3R. (2022). *Diretrizes de auditabilidade e conformidade no desenvolvimento e testes de soluções de IA no âmbito do LIAA-3R / Grupo de Validação Ético-Jurídica (GVEJ) do LIAA-3R, iLabTRF3, iJuspLab* (2. ed., rev. e atual). São Paulo.

Ligot, D. V.. AI governance: A framework for responsible AI development. Available at SSRN: https://ssrn.com/abstract=4817726.

Lin, Y. T., Hung, T. W., & Huang, L. T. L. (2021). Engineering equity: How AI can help reduce the harm of implicit bias. *Philosophy and Technology, 34*(Suppl. 1), 65–90. https://doi.org/10.1007/s13347-020-00406-7

Locher, M. A., & Bolander, B. (2019). Ethics in pragmatics. *Journal of Pragmatics, 145*, 83–90. ISSN 0378–2166 https://doi.org/10.1016/j.pragma.2019.01.011.

Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., & Jacquet, A. (2024). *Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering. Association for Computing Machinery. New York, NY, USA* (Vol. 56, p. 7). ISSN 0360-0300. https://doi.org/10.1145/3626234

Ma, L., Zhang, Z., & Zhang, N. (2018). Ethical dilemma of artificial intelligence and its research progress. *IOP Conference Series: Materials Science and Engineering, 392*, Article 062188. https://doi.org/10.1088/1757-899X/392/6/062188

Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research, 120*, 262–273. https://doi.org/10.1016/j.jbusres.2020.07.045

Makhlouf, K., Zhioua, S., & Palamidessi, C. (2021). Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management, 58*(5), Article 102642. https://doi.org/10.1016/j.ipm.2021.102642

Maluf, S. (1995). *Teoria Geral do Estado* (23ª ed., pp. 205–208). Editora Saraiva. São Paulo.

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. *AI and Ethics, 2*(4), 603–609. https://doi.org/10.1007/s43681-022-00143-x

Manzini, E. J. (2004). *Entrevista semi-estruturada: análise de objetivos e de roteiros. Seminário Internacional sobre Pesquisa e Estudos Qualitativos, 2, 2004, Bauru. A pesquisa qualitativa em debate. Anais*. Bauru: USC. isbn:85-98623-01-6.

Martin, K., & Parmar, B. (). *AI and the Creation of Knowledge Gaps: The ethics of AI transparency*. Available at SSRN: https://ssrn.com/abstract=4207128 https://doi.org/10.2139/ssrn.4207128.

McGraw, G., Bonett, R., Shepardson, V., & Figueroa, H. (2020). The top 10 risks of machine learning security. *Computer, 53*(6), 57–61. https://doi.org/10.1109/MC.2020.2984868

Medaglia, R., Gil-Garcia, J. R., & Pardo, T. A. (2021). Artificial intelligence in government: Taking stock and moving forward. *Social Science Computer Review, 1–18*. https://doi.org/10.1177/08944393211034087

Meijerink, J., & Bondarouk, T. (2018). Uncovering configurations of HRM service provider intellectual capital and worker human capital for creating high HRM service value using fsQCA. *Journal of Business Research, 82*, 31–45. https://doi.org/10.1016/j.jbusres.2017.08.028

Mezgár, I., & Váncza, J. (2022). From ethics to standards – A path via responsible AI to cyber-physical production systems. *Annual Reviews in Control, 53*, 391–404. ISSN 1367-5788. https://doi.org/10.1016/j.arcontrol.2022.04.002

Micheli, M., Ponti, M., Craglia, M., & Berti Suman, A. (2020). Emerging models of data governance in the age of datafication. *Big Data & Society, 7*(2). https://doi.org/10.1177/2053951720948087

Mikalef, P., Conboy, K., Lundström, J. E., & Popovič, A. (2022). Thinking responsibly about responsible AI and 'the dark side' of AI. *European Journal of Information Systems, 31*(3), 257–268. https://doi.org/10.1080/0960085X.2022.2026621

Mikalef, P., Lemmer, K., Schaefer, C., Ylinen, M., Fjørtoft, S. O., Torvatn, H. Y., … Niehaves, B. (2022). Enabling AI capabilities in government agencies: A study of determinants for European municipalities. *Government Information Quarterly, 39*(4), Article 101596.

Ministério da Ciência Tecnologia e Inovação do Brasil. (2021). Estratégia Brasileira de Inteligência Artificial. Governo do Brasil. Retrieved from https://www.gov.br/mcti/pt-br/acompanhe-o-mcti/transformacaodigital/arquivosinteligenciaartificial/ia_estrategia_documento_referencia_4-979_2021.pdf Accessed September 23, 2024.

Ministero dello sviluppo economico. (2019). Proposte per uma Strategia italiana per l'intelligenza artificiale. Retrieved from https://www.mise.gov.it/images/stories/documenti/Proposte_per_una_Strategia_italiana_AI.pdf.

Ministry of Economic Affairs and Employment of Finland. (2017). Suomen tekoälyaika. Retrieved from https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/160391/TEMrap_47_2017_verkkojulkaisu.pdf?sequence=1&isAllowed=y Accessed September 23, 2024.

Ministry of Economic Affairs and Employment of Finland. (2019). *Leading the Way into the Era of Artificial Intelligence: Final Report of Finland's Artificial Intelligence Program 2019. Ministry of Economic Affairs and Employment of Finland* (p. 133). Retrieved from http://urn.fi/URN:ISBN:978-952-327-437-2 Accessed September 23, 2024.

Misuraca, G., & van Noordt, C. (2020). *Overview of the use and impact of AI in public services in the EU, EUR 30255 EN. 2020*. Luxembourg: Publications Office of the European Union. ISBN 978–92–76-19540-5, doi:10.2760/039619, JRC120399.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics, 26*, 2141–2168. https://doi.org/10.1007/s11948–019-00165-5

MRE. (2022). De outros países no Brasil. Retrieved from https://www.gov.br/mre/pt-br/assuntos/Embaixadas-Consulados-Missoes/de-outros-paises-no-brasil Accessed September 23, 2024.

Nagbøl, P. R., & Müller, O. (2020). X-RAI: A framework for the transparent, responsible, and accurate use of machine learning in the public sector. In *IFIP EGOV-ePart-CeDEM conference. p. 259. CEUR workshop proceedings*.

Nagbøl, P. R., Müller, O., & Krancher, O. (2021). Designing a risk assessment tool for artificial intelligence systems. In the next wave of sociotechnical design. In *, 16. 16th international conference on design science research in information systems and technology, DESRIST 2021, Kristiansand, Norway, august 4–6, 2021, proceedings* (pp. 328–339). Springer International Publishing. https://doi.org/10.1007/978-3-030-82405-1_32.

National Institute of Advanced Industrial Science and Technology. (2022). Machine Learning Quality Management Guideline – 2nd English edition. Government of Japan – Digital Architecture Research Center. Retrieved from https://www.digiarc.aist.go.jp/en/publication/aiqm/aiqm-guideline-en-2.1.1.0057-e26-signed.pdf Accessed September 23, 2024.

Navarro, S., Llinares, C., & Garzon, D. (2016). Exploring the relationship between cocreation and satisfaction using QCA. *Journal of Business Research, 69*(4), 1336–1339.

NIST. (2022). AI Risk Management Framework: first draft. Retrieved from https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf Accessed September 23, 2024.

Nordic Council of Ministers. (2018). AI in the Nordic-Baltic region. Retrieved from https://www.norden.org/en/declaration/ai-nordic-baltic-region Accessed September 23, 2024.

Norwegian Data Protection Authority. (2018). Datatilsynet. Retrieved from https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf Accessed September 23, 2024.

Norwegian Ministry of Local Government and Modernisation. (2020). National Strategy for Artificial Intelligence. Retrieved from https://www.regjeringen.no/contentassets/1febbbb2c4fd4b7d92c67ddd353b6ae8/en-gb/pdfs/ki-strategi_en.pdf Accessed September 23, 2024.

Ntoutsi, E., Fafaliosb, P., Gadirajua, U., Iosidisa, V., Nejdla, W., Vidalc, M., … Staab, S. (2020). Bias in Dat-driven AI systems – An introductory survey. *arXiv*. https://doi.org/10.48550/arXiv.2001.09762, 2001.09762v1 [cs.CY].

OECD. (2022a). Artificial Intelligence Observatory. Retrieved from https://oecd.ai/en/ Accessed September 23, 2024.

OECD. (2022b). United States – Local state and Federal Regulations on facial recognition technologies. Organization for Economic and co-operation Development Artificial Intelligence Observatory. Retrieved from retrieved from https://oecd.ai/en/dashboards/policy-initiatives/http:%2F%2Faipo.oecd.org%2F2021-data-policyInitiatives-26890.

OECD. (2022c). Policy Initiatives for Emerging AI-related regulation, Civil society. https://oecd.ai/en/dashboards/policy-initiatives?conceptUris=http:%2F%2Fai.oecd.org%2Fmodel%23Emerging_technology_regulation%7C%7Chttp:%2F%2Fai.oecd.org%2Ftaxonomy%2FtargetGroups%23TG16 Accessed September 23, 2024.

OECD/CAF. (2022). *The strategic and responsible use of artificial intelligence in the public sector of Latin America and the Caribbean*. Paris: OECD Public Governance Reviews, OECD Publishing. https://doi.org/10.1787/1f334543-en

Ojo, A., Mellouli, S., & Ahmadi Zeleti, F. (2019). A realist perspective on AI-era public management. In *In 20th annual international conference on digital government research* (pp. 159–170). ACM.

Oneto, L., & Chiappa, S. (2020). *Fairness in machine learning. arXiv:2012.15816 [cs.LG]*.

Özdemir, V., & Hekim, N. (2018). Birth of industry 5.0: Making sense of big data with artificial intelligence,"the internet of things" and next-generation technology policy. *OMICS: A Journal of Integrative Biology, 22*(1), 65–76. https://doi.org/10.1089/omi.2017.0194

Papagiannidis, E., Enholm, I. M., Dremel, C., Mikalef, P., & Krogstie, J. (2023). Toward AI governance: Identifying best practices and potential barriers and outcomes. *Information Systems Frontiers, 25*, 123–141. https://doi.org/10.1007/s10796-022-10251-y

Phillips, P. J., Hanan, C. A., Fontana, P. C., Broniatowski, D. A., & Przybocki, M. A. (2021). *Four principles of explainable artificial intelligence*. National Institute of Standards and Technology. U.S. Department of Commerce.

Pinski, M., Hofmann, T., & Benlian, A. (2024). AI literacy for the top management: An upper echelons perspective on corporate AI orientation and implementation ability. *Electronic Markets, 34*, 24. https://doi.org/10.1007/s12525-024-00707-1

Presidencia de la Nación. (2019). ARGENIA – Plan Nacional de Inteligencia Artificial. Argentina. Retrieved from https://oecd-opsi.org/wp-content/uploads/2021/02/Argentina-National-AI-Strategy.pdf Accessed September 23, 2024.

Ragin, C. C. (2008). *Redesigning Social Inquiry: Fuzzy Sets and Beyond* (pp. 85–97). Chicago: Univ. of Chicago Press.

Rahwan, I. (2017). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology, Springer., 20*, 5–14. https://doi.org/10.1007/s10676-017-9430-8

Rahwan, I., Cebrian, M., Obradovich, N., et al. (2019). Machine behaviour. *Nature, 568*, 477–486. https://doi.org/10.1038/s41586-019-1138-y

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine, 169*(12), 866–872. https://doi.org/10.7326/M18-1990

Rhahla, M., Allegue, S., & Abdellatif, T. (2021). Guidelines for GDPR compliance in big data systems. *Journal of Information Security and Applications, 61*, Article 102896. https://doi.org/10.1016/j.jisa.2021.102896

Rihoux, B., & Ragin, C. (2008). *Configurational comparative methods: Qualitative comparative analysis (QCA) and related techniques*. London and Thousand Oaks, CA: Sage.

Roorda, S. A. H. (2021). *Facial recognition for public safety a supportive tool for the municipal decision-making process on using facial recognition for public safety, the FRPS risk governance method*. Master's thesis,. University of Twente.

Rose, J., Flack, L. S., & Sæbø, Ø. (2018). Stakeholder theory for the E-government context: Framing a value-oriented normative core. *Government Information Quarterly, 35*(3), 362–374. ISSN 0740-624X https://doi.org/10.1016/j.giq.2018.06.005.

Roselli, D., Matthews, J., & Talagala, N. (2019). *Managing Bias in AI. In Companion Proceedings of The 2019 World wide web conference (WWW '19)* (pp. 539–544). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3308560.3317590

Ruijer, E. (2021). Designing and implementing data collaboratives: A governance perspective. *Government Information Quarterly, 38*(4), Article 101612. https://doi.org/10.1016/j.giq.2021.101612

Saldaña, J. (2013). *The coding manual for qualitative researchers* (2nd ed.). Sage.

Schaefer, C., Lemmer, K., Samy Kret, K., Ylinen, M., Mikalef, P., & Niehaves, B. (2021). Truth or dare?–how can we influence the adoption of artificial intelligence in municipalities?. Retrieved from http://hdl.handle.net/10125/70899.

Schrader, D., & Ghosh, D. (2018). Proactively protecting against the singularity: Ethical decision making AI. *IEEE Computer and Reliability Societies Review, 16*(3), 56–63.

Schüller, K. (2022). Data and AI literacy for everyone. *Jan, 1*, 477–490. https://doi.org/10.3233/sji-220941

Shao, S., Zhao, R., Yuan, S., Dig, M., & Wang, Y. (2022). Tracing the evolution of AI in the past decade and forecasting the emerging trends. *Expert Systems with Applications., 209*, Article 118221. https://doi.org/10.1016/j.eswa.2022.118221

Sharma, G. D., Yadav, A., & Chopra, R. (2020). Artificial intelligence and effective governance: A review, critique and research agenda. *Sustainable Futures, 2*. https://doi.org/10.1016/j.sftr.2019.100004. ISSN 2666–1888.

Sigfrids, A., Leikas, J., Salo-Pöntinen, H., & Koskimies, E. (2023). Human-centricity in AI governance: A systemic approach. *Frontiers in Artificial Intelligence, 6*, Article 976887. https://doi.org/10.3389/frai.2023.976887. PMID: 36872934; PMCID: PMC9979257

Silberg, J., & Manyika, J. (2019). *Notes from the AI frontier: Tackling bias in AI (and in humans). 1(6)*. McKinsey Global Institute. Retrieved from https://www.mckinsey.com/~/media/mckinsey/featured%20insights/artificial%20intelligence/tackling%20bias%20in%20artificial%20intelligence%20and%20in%20humans/mgi-tackling-bias-in-ai-june-2019.pdf Accessed September 23, 2024.

Silveira, M. B., Saldanha, R. P., Leite, J. C. C., Silva, T. O. F. D., Silva, T., & Filippin, L. I. (2018). Construction and validation of content of one instrument to assess falls in the elderly. *einstein (Sao Paulo), 11;16*(2), Article eAO4154. Retrieved from https://doi.org/10.1590/S1679-45082018AO4154 accessed September 23, 2024.

Sivarajah, U., Kamal, M. M., Irani, Z., & Weerakkody, V. (2017). Critical analysis of big data challenges and analytical methods. *Journal of Business Research, 70*, 263–286.

Smuha, N. A. (2021). Beyond a Human Rights-Based Approach to AI Governance: Promise, Pitfalls, Plea. *Philos. Technol., 34*(Suppl 1), 91–104. https://doi.org/10.1007/s13347-020-00403-w

Stahl, B. C., Antoniou, J., Ryan, M., et al. (2022). Organisational responses to the ethical issues of artificial intelligence. *AI & SOCIETY, 37*, 23–37. https://doi.org/10.1007/s00146-021-01148-6

Stix, C. (2021). Foundations for the future: Institution building for the purpose of artificial intelligence governance. *AI and Ethics*. https://doi.org/10.1007/s43681-021-00093-w

Strauß, S. (2021). Deep automation Bias: How to tackle a wicked problem of AI? *Big Data Cogn. Comput., 5*, 18. https://doi.org/10.3390/bdcc5020018

Switzerland Federal Council. (2021). Guidelines on Artificial Intelligence for the Confederation. Retrieved from https://www.sbfi.admin.ch/dam/sbfi/en/dokumente/2021/05/leitlinien-ki.pdf.download.pdf/leitlinien-ki_e.pdf Accessed September 23, 2024.

Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society, 40*(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377

Tangi, L., van Noordt, C., Combetto, M., Gattwinkel, D., & Pignatelli, F. (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. ISBN 978-92-76-53058-9, doi:10.2760/39336, JRC129301.

Torlig, E. G. S., Resende-Júnior, P. C., & Fujihara, R. K. (2019). Proposição de uma Nova Orientação para Validação de Roteiros em Pesquisas Qualitativas. In *XLIII Encontro da ANPAD – EnANPAD 2019, São Paulo*.

Torlig, E. G. S., Resende-Júnior, P. C., Fujihara, R. K., Demo, G., & Montezano, L. (2022). Validation Proposal for Qualitative Research Scripts (Vali-Quali). *Administração: Ensino e Pesquisa, 23*(1). https://doi.org/10.13058/raep.2022.v23n1.2022, 4-29; Jan-Abr.

Vero. (2019). Finnish Tax Administration's ethical principles for AI. Retrieved from https://www.vero.fi/en/About-us/finnish-tax-administration/operations/responsibility/finnish-tax-administrations-ethical-principles-for-ai/#:~:text=Our%20AI%20follows%20laws%20and%20regulations&text=The%20use%20of%20AI%20does,our%20partners%20carefully%20and%20responsibly.

Vetrò, A., Torchiano, M., & Mecati, M. (2021). A data quality approach to the identification of discrimination risk in automated decision-making systems. *Government Information Quarterly, 38*(4). https://doi.org/10.1016/j.giq.2021.101619. ISSN 0740-624X.

Vial, G. (2019). Understanding digital transformation: A review and a research agenda. *The Journal of Strategic Information Systems, 28*(2), 118–144. ISSN 0963-8687. https://doi.org/10.1016/j.jsis.2019.01.003

Vilminko-Heikkinen, R., & Pekkola, S. (2019). Changes in roles, responsibilities and ownership in organizing master data management. *International Journal of Information Management, 47*, 76–87. ISSN 0268-4012, https://doi.org/10.1016/j.ijinfomgt.2018.12.017.

Vining, R., McDonald, N., McKenna, L., Ward, M. E., Doyle, B., Liang, J., … Fogarty & Brennan, R. (2022). Developing a Framework for Trustworthy AI-Supported Knowledge Management in the Governance of Risk and Change. In *HCI International 2022-Late Breaking Papers. Design, User Experience and Interaction: 24th International Conference on Human-Computer Interaction, HCII 2022, Virtual Event, June 26–July 1, 2022, Proceedings* (pp. 318–333). Cham: Springer International Publishing.

WEF. (2020). Unblocking Public Sector AI – AI procurement in a Box: Project overview. Retrieved from https://mkto.deloitte.com/rs/712-CNF-326/images/Retningslinjer-for-offentlige-AI-anskaffelser.pdf Accessed September 23, 2024.

WEF. (2020b). Unblocking Public Sector AI. AI Procurement in a Box: Pilot case studies from the United Kingdom. Retrieved from https://www.weforum.org/reports/ai-procurement-in-a-box/#case-study-uk Accessed September 23, 2024.

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2018). Artificial intelligence and the public sector—Applications and challenges. *International Journal of Public Administration, 42* (7), 596–615.

Wirtz, B. W., Weyerer, J. C., & Kehl, I. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government Information Quarterly*. https://doi.org/10.1016/j.giq.2022.101685. ISSN 0740-624X.

Wright, S. A., & Schultz, A. (2018). The rising tide of artificial intelligence and business automation: Developing an ethical framework. *Business Horizons, 61*(6), 823–832. https://doi.org/10.1016/j.bushor.2018.07.001

Xue, M., Yuan, C., Wu, H., Zhang, Y., & Liu, W. (2020). Machine learning security: Threats, countermeasures, and evaluations. *IEEE Access, 8*(74720–74742), 2020. https://doi.org/10.1109/ACCESS.2020.2987435

Zhou, J., & Chen, F. (2023). AI ethics: From principles to practice. *AI & SOCIETY, 38*, 2693–2703. https://doi.org/10.1007/s00146-022-01602-z

Zicari, R. V., Brodersen, J., Brusseau, J., Düdder, B., Eichhorn, T., Ivanov, T., … Westerlund, M. (2021). Z-inspection®: A process to assess trustworthy AI. *IEEE Transactions on Technology and Society, 2*(2), 83–97. https://doi.org/10.1109/TTS.2021.3066209

Zick, T., Kortz, M., Eaves, D., & Doschi-Velez, F. (2024). *AI Procurement Checklists: Revisiting Implementation in the Age of AI Governance*. arXiv:2404.14660v1 [cs.CY]. https://doi.org/10.48550/arXiv.2404.14660.

Zuiderwijk, A., Chen, Y., & 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). https://doi.org/10.1016/j.giq.2021.101577. ISSN 0740-624X.

Patricia Gomes Rêgo de Almeida has Master degree at University of Rio Grande do Norte, Department of Electrical Engineering, and PhD degree at the University of Brasilia (UnB), Department of Business Administration, where she has developed research regarding artificial intelligence regulation (hard and soft law) and artificial intelligence governance. She is member of Linselab at University of Brasilia (UnB). She is the coordinator of Digital Innovation, Governance and Strategy at the Brazilian chamber of Deputies, where she is responsible for its Digital Transformation Strategy, Artificial Intelligence Governance and Data Fluency Program. She is coordinator of the Parliamentary Data Science Hub at the Inter-parliamentary Union, where she coordinated the writing of Guidelines for AI in Parliaments.

Carlos Denner dos Santos Júnior Junior is professor at University of Brasilia (UnB), Department of Business Administration, focusing on the intersection between administration and computing, developing cutting-edge scientific knowledge about the creation and management of new businesses and information technologies with applications in public administration and their impacts on the market in general. He has post-doctoral degrees at: Université du Quebec à Montréal (UQUAM), Universidade Federal de Pernambuco (UFPE), University of Nottingham, and Universidade de São Paulo (USP). He is the coordinator of the Research Group at CNPq Sociedados - on the Strategic and Competitive Use of Data (Open) and Software (Free), and the DINTER coordinator between UnB-UNIMONTES.

Artificial Intelligence Regulation: a framework for governance

Article *in* Ethics and Information Technology · September 2021

DOI: 10.1007/s10676-021-09593-z

Patricia Gomes Rêgo de Almeida
Chamber of Deputies (Brazil)

Carlos Denner dos Santos Jr.
University of Brasília

Josivania Silva Farias
University of Brasília

Ethics and Information Technology
https://doi.org/10.1007/s10676-021-09593-z

ORIGINAL PAPER

Artificial Intelligence Regulation: a framework for governance

Patricia Gomes Rêgo de Almeida[1,2] · Carlos Denner dos Santos[1,3] · Josivania Silva Farias[1]

Accepted: 7 April 2021

Abstract

This article develops a conceptual framework for regulating Artificial Intelligence (AI) that encompasses all stages of modern public policy-making, from the basics to a sustainable governance. Based on a vast systematic review of the literature on Artificial Intelligence Regulation (AIR) published between 2010 and 2020, a dispersed body of knowledge loosely centred around the "framework" concept was organised, described, and pictured for better understanding. The resulting integrative framework encapsulates 21 prior depictions of the policy-making process, aiming to achieve gold-standard societal values, such as fairness, freedom and long-term sustainability. This challenge of integrating the AIR literature was matched by the identification of a structural common ground among different approaches. The AIR framework results from an effort to identify and later analytically deduce synthetic, and generic tool for a country-specific, stakeholder-aware analysis of AIR matters. Theories and principles as diverse as Agile and Ethics were combined in the "AIR framework", which provides a conceptual lens for societies to think collectively and make informed policy decisions related to what, when, and how the uses and applications of AI should be regulated. Moreover, the AIR framework serves as a theoretically sound starting point for endeavours related to AI regulation, from legislation to research and development. As we know, the (potential) impacts of AI on society are immense, and therefore the discourses, social negotiations, and applications of this technology should be guided by common grounds based on contemporary governance techniques, and social values legitimated via dialogue and scientific research.

Keywords Ethics · Artificial Intelligence · Regulation · Governance · Framework

Introduction

The widespread use of AI in our daily actions and in an unnoticeable fashion (Cerka et al., 2015) has introduced unprecedented ethical issues to a broad and complex social system (Cave et al., 2019).

From the same perspective, the complexity of data treatment in the design and development process of a machine learning solution increases the likelihood of ethical surprises, which demands a wider evaluation of the ethical and social impacts (Butterworth, 2018).

Based on this reflection, this work has sought to conduct a vast search for literature that is relevant in terms of Artificial Intelligence Regulation, processing and grouping it into a set of purposes presented as frameworks or guidelines for a framework based on ethical principles. Their main contributions have been customised as a framework based on the Design and Action Theory (Gregor, 2006) that allows for reflections and actions aimed at regulating and governing operations and relationships between natural and legal persons on one side, and AI-embedded systems on the other.

✉ Patricia Gomes Rêgo de Almeida
patricia.almeida@camara.leg.br

Carlos Denner dos Santos
carlosdenner@unb.br

Josivania Silva Farias
josivania@unb.br

[1] University of Brasilia (UnB) – Department of Administration, Brasília, Brazil

[2] Chamber of Deputies of Brazil – Directorate of Innovation and Information Technology, Brasília, Brazil

[3] LATECE, University of Quebec at Montreal (UQAM), Montreal, Canada

Published online: 21 April 2021

Reasons to regulate AI

Since the term was coined in 1956, Artificial Intelligence has been associated with a wide range of concepts (Cerka et al., 2017; Jackson, 2019) based on a thinking human being and on rational behaviour, which could be synthetised as: systems that think and act like humans and systems that think and act rationally (Cerka et al., 2015; Russell & Norvig, 1995). Equally wide is the variety of different names associated with whatever utilises AI technology: robots, smart systems, intelligent systems, intelligent agents, AI agents, AI algorithms, intelligent algorithms, and autonomous systems, to mention a few.

For the purpose of avoiding misunderstandings regarding AI, the High-Level Expert Group established by the European Commission has defined AI systems as “software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information, derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.” (AI HLEG, 2019a). Considering the difficulty of defining AI in a way that could fit all approaches needed for regulation and governance actions with clear communication among all stakeholders, this article adopts the definition of AI established by the High-Level Expert Group on AI systems.

The responsibilities, security, intellectual property, and privacy associated with different systems for medical robots, drones, autonomous cars, among several "intelligent solutions" offered every day have been questioned.

Illustrating the level of risk-related indeterminacy, machine learning has been combined with game theory (Conitzer et al., 2017) in cases where developers were using game theory to help teach strategic defence to algorithms. A game between two algorithms predicted that one would kill the other only in case of an absolute scarcity of resources. However, when a more intelligent algorithm was introduced, it immediately killed the weaker ones (Firth-Butterfield, 2017). This case reinforces the idea that an autonomous system will inevitably find itself in a situation in which it needs not only to obey a certain rule or not, but also to make a complex ethical decision (Dennis et al., 2016).

Facing the risks compels us to explore their causes and effects. Although the effects of AI are not yet known, a large amount of them can currently be classified. Firstly, those coming from the undesired effects, such as biases, discrimination, loss of privacy, false positives and false negatives, loss of autonomy, (psychological, financial, or physical) damage, loss of control, difficulty identifying liabilities, losses or decreases in human rights, unemployment, misjudgements, and concentration of power and wealth in a few companies. Secondly, some risks are the result of intentional misuses, such as fake news, deep fakes, cyberattacks, terrorism, warfare, weapons, people manipulation, espionage, low level of democracy (Beltran, 2020; Benjamins & Garcia 2020; Borgesius, 2018; Jackson, 2020; Jobin et al., 2019; Mika et al., 2019).

Considering all those risks, establishing best practices for delegating and defining new moral responsibility attribution models is crucial to leverage the opportunities created by AI (Taddeo & Floridi, 2018). Risk assessment models can provide support and flexibility for Big Data and AI applications (Mantelero, 2018), and stakeholders who develop and deploy AI-based systems must enhance their knowledge of the values protected by human rights and how those rights apply to their own actions (Smuha, 2020).

Despite being a huge challenge, finding a way to deal with ethical issues must be a constant target of research, for what we need to join all our forces (Bostrom, 2014), and AI regulation is on the right path to get there (Carter, 2020).

The reasons to regulate include: manufacturers’ need to comprehend a legal framework within which they can operate reliably; consumers’ and society’s need to be protected from devices that may harm or adversely affect them; and the need for business opportunities (Holder et al., 2016a).

In industries still lacking regulation, the general approach observed is that innovation is freely allowed, but those in charge should bear the consequences in case certain types of damage are caused (Reed, 2018).

Faced with the challenge of minimising those risks, a combination of strategy and actions must be put to practice during the entire lifecycle of AI systems, in order not only to identify damages and responsibilities, but also, and especially, to avoid them.

Seeking the best way to regulate

Sometimes, when used to denote an attempt to standardise behavioural patterns, the term “regulation” assumes the meaning of a law (Hildebrandt, 2018).

However, on a broader approach, regulation is a sustained attempt to modify behaviours of others according to defined standards or purposes in order to produce the desired outcomes. This can involve standard-setting, information-gathering, and behaviour modification mechanisms (Black, 2002), especially in cases evolving ethical issues, whose understanding is complex when applied to a real world. Therefore, law is just one way of regulating society, while other alternatives to regulate human behaviour may also be widely used (Hildebrandt, 2018).

Disruptive innovation always challenges regulatory strategies due to the reactive nature of traditional regulation (Kaal & Vermeulen, 2017). In the case of innovation by AI, the challenge is amplified, since it is strongly related to ethical issues and its results could be unpredictable in some situations, bringing about unforeseen social impacts. In addition, if AI adoption and implementation are conducted in a reckless manner, social and political instability could ensue, thus threatening freedom, self-determination, human rights, and fundamental values (Caron & Gupta, 2020). As human behaviour encompasses decisions from an ethical perspective, the regulation should also consider it. While norms as instruments of regulation relate to what is good or bad from society's point of view, ethics concerns itself with the nature of the principles upon which those norms are founded (Pedro, 2014).

A few laws have been resorted to in an attempt to settle damages caused by AI-supported products and services judicially. If, on the one hand, the number of cases is multiplying, on the other, the legislative branch seems to be moving at a negligible speed compared to the technological advancements enforcing the perception that traditional regulation does not fit in this challenge (Cerka et al., 2015; Larsson, 2020; Villaronga & Heldeweg, 2018). Part of this increasing gap between laws and technology is caused by the lack of a thorough and accurate definition of AI (Firth-Butterfield, 2017; Larsson, 2020), which is aggravated by the fact that the definition changes as the technology evolves (Fjeld et al., 2020). Considering this issue, the concept of dynamic regulation could fit in the field of AI, as it is based on learning by doing and continuity of regulatory relationships (Kaal & Vermeulen, 2017; Lewis & Yildirim, 2002).

A yet-to-be-solved equation is the breadth of laws dealing with globally produced and commercialised technologies (Holder et al., 2016a) and robot-generated inventions (Holder et al., 2016b). The problem reaches even broader dimensions when one considers the complex networks established in the technology industry, making it possible for products to be subjected to learning from data scattered across the world (Lenardon, 2017).

Large-scale data analyses have revealed that the key challenge related to the AI regulation dilemma is demonstrating it is produced and deployed appropriately (Butterworth, 2018). One of the most advocated strategies is transparency, an opening of the entire production process, especially the decision-making rules, the method, and the data utilised when training the intelligent system (Buiten, 2019; Butterworth, 2018; Tutt, 2017). However, on certain occasions, even in case the AI algorithm is open, full transparency cannot be ensured, as there is a difference between seeing the whole code and understanding all of its potential effects (Firth-Butterfield, 2017). A similar strategy to open data is the Explainable Artificial Intelligence (XAI) standard for the creation of coding models oriented towards a global comprehension (Adadi & Berrada, 2018; Taddeo & Floridi, 2018). In addition to the concerns related to the development process of an AI system, data governance has been recognised as being key to AI governance (Hilb, 2020; UK Government, 2018).

Some of the AI regulation theories that have been proposed are based on contractual and extracontractual liability, or on strict liability, and adopt a liability model in which the moral responsibility is distributed among designers, regulators, and users. The attempt to hold robots accountable for their actions has led a few countries to consider the possibility of granting a legal identity to each unit. One could argue that if parties in a contractual relationship may be legally represented by another entity, then so can systems (Cerka et al., 2017). As a counterargument, the term "robot liability" should be replaced with "indirect liability over the robot", given the impossibility of claiming damages from a robot, i.e., it cannot be held criminally liable. Thus, the impact of such products on society should also be a liability (Jackson, 2019; Nevejans, 2016). Although this latter understanding tends to be more acceptable from a global perspective, a liability model is still an essential and complex variable to be defined through an AI regulation strategy.

Also among the concerns that motivate AI regulation is the approach aimed at minimising the disruption of the work model with the goal of fighting job loss (Wright & Schultz, 2018).

Drawing attention to the domain of what is to be regulated, attempts to legislate digital technologies without proper knowledge for doing so have been criticised (Reed, 2018). With the intention of minimising those risks, a gradual regulation strategy (Villaronga & Heldeweg, 2018) can be used. When mitigating risks, regulatory agencies could bar the introduction of certain algorithms into the market until their safety and efficacy have been proven by means of tests (Tutt, 2017) founded on ethics (Arkin, 2011).

In 2017, the European Parliament Committee on Legal Affairs released a report recommending the creation of a European agency for robotics and AI, suggesting a combination of both hard and soft laws, given the complexity associated with the evolution of the regulatory model. It would put regulators and external experts together to monitor AI trends and study standards for best practices (Cath et al., 2017; Nevejans, 2016). After approving the study of the High-Level Expert Group on AI, the European Commission recommended upgrading the European Framework to one especially designed for AI Governance (European Commission, 2019). In the same direction, the House of Lords (2018) has recommended the creation of an AI regulatory framework.

Another effort observed in the US has resulted in S.3891, which defines conditions for advancing Artificial Intelligence research, including the development of technical standards (US Congress, 2020), and in H.Res. 153, which aims to support the development of guidelines for the ethical dsevelopment of Artificial Intelligence (US Congress, 2019).

In a parallel effort, many self-regulatory private-sector initiatives have been created, and research has been carried out to discuss ethical issues on AI development and use, such as the Partnership on AI to Benefit People and Society (AI4People, 2018; Partnership on AI, 2016; The Future of Life Institute, 2019b), The Montreal Declaration for a Responsible Development of Artificial Intelligence (University of Montreal, 2018), and The Toronto Declaration (Toronto, 2020).

At the government level, ethical principles were considered when the national AI-oriented strategies of a few countries were drawn up, as happened in Japan (Japanese Cabinet Office, 2019), France (French PM, 2018), Germany (German Federal Government, 2018), United Arab Emirates (Dubai, 2019), India (Aayog, 2018), and Singapore (Monetary Authority of Singapore, 2019). Additionally, several countries have shown their intention to create policies and laws to regulate the development and use of AI (Future of Life Institute, 2019a). Similar concerns have served as the basis for recommendations regarding ethical principles by a few transnational organisations, such as the Council of Europe (2018) and the Organisation for Economic Cooperation and Development (2019).

As the major concern regarding both self-regulation and government initiatives kickstarted the debate on AI governance through ethical principles, a set of core topics was comprised in each one of them: privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values. However, different principles can be observed under the same topic, which illustrates the lack of unanimity (Fjeld et al., 2020).

Standardisation documents are also part of the efforts associated with the AI regulation challenge. A good example is the ad hoc technical committee on Artificial Intelligence established within the International Organisation for Standardisation (ISO), whose plan includes two dozen standards on AI and Big Data (Neznamov, 2020).

The huge gap between ethical guidelines and laws, apart from the great number of potential situations in which stakeholders (developers, deployers, etc.) fail to apply such ethical principles, draws attention to the need to shift from principles to processes when it comes to AI governance (Larsson, 2020). Thus, there is a long path to be paved through a connected network of processes including several stakeholders in a way to keep on pace with the society's values.

Method

With the goal of surveying the relevant scientific literature on AI regulation, we have systematically searched for and organised papers to summarise the corpus and perform a qualitative analysis to understand the evolution and current state of the science.

We have compiled papers published between 2010 and 2020 containing the following expressions: ("ARTIFICIAL INTELLIGENCE" and "ETHICAL USE"), ("ARTIFICIAL INTELLIGENCE" and "REGULATION"), or ("ARTIFICIAL INTELLIGENCE" and "GOVERNANCE"). This search then resulted in title and subject matches on the ScienceDirect, JSTOR, SpringerLink, PROQUEST, IEEE, Scopus, DOAJ, and Google Scholar databases. Only peer-reviewed research articles in English have been compiled.

The selection of papers was later refined by reading all abstracts with the goal of removing case-specific discussions, as well as those in which regulation was not the main topic under debate. In addition, new papers on AI-related laws and government strategies that presented arguments on ethical principles have also been included in the sample.

This final corpus of literature has been classified according to specific parameters: year of publication, journal, author, author's institution, author's field of study, country, keywords. Summaries of each paper have also been developed to include: concepts, findings, contributions, agenda, approach, method, and researched subject. The following terms were considered when classifying the articles: "ethics/ethical principles", "how to regulate/existing regulation", "government strategies", and "framework or guidelines similar to a framework". After analysing the abstracts, a sample comprising 109 documents was selected for further reading and inclusion.

Results

In chronological terms, it is worth highlighting that 88.1% of the papers were published after 2015, with a growing production every year following that.

The sample reflects the evolution in the fields of research that take an interest in AI regulation, which is desired (Floridi et al., 2020). Although Artificial Intelligence as a subject of study traditionally pertains to Information Technology (Computer Science and Engineering), there has been a growing interest in its regulation by other areas, such as Law, Business Administration, and Philosophy. Out of the entire sample, researchers from the field of IT (combined or not with other fields) represent 47.7% and researchers exclusively from IT represent 27.5%, whereas researchers from other fields (excluding IT) represent 52.3%. In some cases, the same article is co-authored by researchers from different areas (22%).

Special heed has been paid to the analysis of the main object of the sample, non-exclusively divided into: "ethics and ethical principles" (45.9%), "how to regulate and existing regulation" (47.7%), "government strategies" (9.2%) and "framework or guidelines similar to a framework" (19.3%). It is worth noting that discussions on how to regulate only became significant in 2016. Concerning the discussions on AI regulatory frameworks, AI guidelines based on ethical principles with orientations similar to a framework have also been considered when they go beyond a description of ethical principles and actually provide orientations concerning how to apply those principles. From that perspective, 21 unique models have been found, which will be presented and analysed below.

Model for ethical issues in experimental technologies (Amigoni & Schiaffonati, 2018)

Based on the premise that a robot is an experimental technology, this model intends to minimise the ethical dilemmas associated with decisions made by autonomous systems (Poel, 2016). The proposal supports decision-making processes based on 16 conditions for deploying experimental technologies built to anticipate potential ethical issues as robots interact with people and the environment. Split into three groups, the conditions are aimed at: preventing damages (non-maleficence conditions: means for gaining knowledge of risks and benefits, monitoring of data and risks, possibility and willingness to adapt or terminate the experiment, risk mitigation, consciously scaling up, flexible setup, and avoidance of locking-in and undermining resilience); good-doing (beneficence conditions: expectation of social benefits, clear distribution of responsibilities); and respect for autonomy and justice (experimental subjects informed, approval by democratically legitimised bodies, possibility of experimental subjects influencing the project, possibility of withdrawing subjects from the experiment, special treatment given to vulnerable experimental subjects, fair distribution of potential hazards and benefits, reversibility of compensation of harm).

The model is an approach to regulation through a development process that would be part of a gradual interactive strategy set forth during the design stage. One can find among the outputs the epistemological role of exploratory experiments, while acquiring the knowledge of how robots behave in a real-world scenario. The authors highlight the prediction of "red button" conditions for situations in which the risk of harming people cannot be securely avoided during the experiment.

The 16 conditions proposed by Amigoni and Schiaffonati's "Ethical Framework for Robot Systems" seem to fit perfectly in standardised processes built by regulatory agencies as they test all the technologies submitted by the industry and service providers. The proposal can also be incorporated through risk analyses conducted by scholars for society as a whole.

Interactive regulatory governance model (Villaronga & Heldeweg, 2018)

Considering that regulatory actions cannot keep up with the speed of technology, and that top-down regulation approaches require mature laws, the authors have identified the need for a hybrid approach to start regulating AI technologies. They argue that bottom-up mechanisms can help develop the legislation and produce knowledge of AI development processes.

Focusing on a balance between regulation/legislation-in-progress and technology-in-progress, the proposal is based on an interactive governance model for technological development and law formulation processes in which the attributions of stakeholders are highlighted through process descriptions. The need for continuous learning and a gradual evolution of the legal framework is noteworthy, using such expressions as "Regulatory Innovation" and "Temporary Experimental Legislation", and considering the proper sequence of actions among agents at the maturity stage of an innovation's lifecycle.

The proposed model includes components such as:

- A Regulatory-to-Technology (R2T) macro-process to guide the creation of a new conceptual model for robots in accordance with the existing legislation, considering how it affects the way intelligent systems are built and used. It enables the creation of an AI technology impact assessment encompassing ethical, legal, and societal consequences. It focuses on legal opportunities or constraints that could have an impact on a new or existing robot. The result of the analysis considers a range of alternatives, from "abort development", "adjust plans", "go-ahead and lobby for legal change", or "take risks".
- A Technology-to-Regulatory (T2R) macro-process to adjust the law to the needs that result from the evolution of technology or the relationship between intelligent systems and society. It allows for the implementation of a regulatory impact assessment.
- A Governance Committee to rule on the reports related to the impact of both R2T (*ex-ante* robot) and T2R (*ex-post* robot) processes.

- A data repository shared by R2T and T2R in order to gather data about whether each AI technology (planned or in use) is in compliance with the law.

Among the main benefits of this hybrid AI Governance Model, it is worth highlighting the integration of top-down and bottom-up regulatory actions in an incremental strategy, thus minimising the risk posed when regulating a new, constantly changing object.

The proposed Interactive Regulatory Governance Model helps to raise awareness regarding the lack of a continuous resource to connect both worlds—technology and legislation—while being iteratively developed and improved. Since the legislative branch is in charge of the legislation (in most democratic countries), it can be associated with the R-side of processes. When looking for the most ideal entity to act as the T-side of processes, the tasks of a regulatory agency can be identified.

Connecting both sides, R2T and T2R processes would be a strategy to establish a closer relationship between the legislative branch and the regulatory agency.

Ethics model for AI development and deployment (Schrader & Ghosh, 2018)

Founded upon philosophical principles and the dimensions associated with safeguarding human rights and well-being, the proposed ethical framework for AI development and deployment has been designed to implement core functions to represent ethical activities and the outcomes from both the philosophical and ethical perspectives.

The ethical perspectives are split into six categories: Rights (deontological ethics); Damages and Goods (teleological ethics); Virtue (aretaic ethics); Community (community ethics); Dialogue (communication ethics); and Flourishing (flourishing ethics).

The recommended core functions to be considered when developing AI systems are:

- Identifying ethical issues of AI—fairness, transparency, equity, goodness, beneficence, social utility, happiness, and protection of humans.
- Raising human awareness of AI—a clear understanding of how AI systems work within each product and how the industry develops algorithms.
- Collaborating with AI—dialogical interaction, listening, and understanding between humans and AI.
- Accountability of AI—guaranteeing the ethical compliance of AI systems and their designers.
- Integrity of AI—maintaining the AI system limited to the purpose for which the technology was intended.

A matrix combining the five core functions with the six perspectives has been built as a guideline to be followed during the AI project. As a proactive action in the design, development, and use of products and services that utilise AI, the model seeks to reflect the nature of social changes demanded by a new ethical thought.

Although they do not associate the framework with any specific organisation or institution, the authors' contribution can be applied by a regulatory agency when auditing the industry, as well as in its internal processes, to better understand the impacts technology has on the stakeholders.

Competency-based AI regulation model (Scherer, 2016)

Considering the competencies, strengths, and weaknesses of each state power, the proposal of an AI Regulatory Model (AIDA—Artificial Intelligence Development Act) is based on the distribution of responsibilities without losing sight of the mission goals. The model acknowledges the regulatory role of the executive, legislative, and judicial powers as agents in the regulatory process.

In the proposed model, the legislative branch would provide a statute placing a regulatory agency in charge of certifying AI products and services with regard to user and social safety. In general, legislators have limited knowledge of AI systems, their only support being a few committee meetings with experts. In order to solve this problem, legislators would delegate the responsibility for policy-making to the regulatory agency.

Supported by groups of researchers, the regulatory agency would comprise two main areas: policy-making and certification. Such an agency would be expected to be more agile and competent to monitor the evolution of technology, identify risks in the intelligent learning process and use of AI, issue technical recommendations, and verify that the technology is being applied for its intended purposes. A certificate would be given to designers, manufacturers, and service providers after being approved through the agency's processes. Pre-certification rules would also be made public to the industry and service providers. In case of an accident with certified products, the agency would publish a report to society, explaining the circumstances behind its occurrence and which certification rules/processes would therefore be modified.

Due to their *ex-post* nature, courts would judge cases considering whether or not a certification exists. Courts would judge companies for any losses and damages caused, considering the situation in which those organisations find themselves when it comes to certification. If a company's products or services cause any damages, if certified, the company would be judged based on more lenient rules, whereas uncertified companies would be subjected to more rigid norms.

The proposed model takes into consideration the natural attributions of each entity within the government. Agility is required for the actions performed by the regulatory agencies, which would give them a prominent role in the regulation process. This is key to enable the evolution of technology while the legislation takes its time to mature.

Regulation model sustained by society (Rahwan, 2017)

Inspired by the Social Contract Theory (Rousseau, 2016), the Regulatory Model Sustained by Society adjusts the "human-in-the-loop (HITL)" to the "society-in-the-loop (SITL)" model.

The use of HITL thinking in AI has been largely applied to help an algorithm learn from humans' contributions. The agility and effectiveness of a HITL interactive learning machine stem from user feedback, thus enriching the knowledge that gets generated.

From a regulation perspective, the author argues that it is not sufficient to only adjust HITL to use a human to monitor an AI system and correct it in case of misbehaviour. By doing so, the regulation would rely on the judgment of an individual or group of individuals that subject the whole process to a narrow analysis. If we want to deal with a system that has an impact on the values of an entire society, that society must be included in the analysis, giving it a broader approach. It would not only avoid biased judgments, but also balance the competing interests of different stakeholders.

It is suggested that SITL be used in a process characterised by human-based government and citizen channels. On one side, the government's AI products and services would be run and, on the other side, citizens would evaluate those smart systems based on their own values. This would allow the government to understand how social behaviour and values change. Therefore, society-in-the-loop would become a governance tool for society to control and proactively identify those elements. Conflicts among safety-, privacy-, and justice-related concepts would benefit from this model. This relationship can be summed up as: society-in-the-loop = human-in-the-loop + social contract. The model also recommends auditing mechanisms to tackle the possibility of fake data manipulated by social groups at the learning stage, as well as results that would affect regulations.

For the purpose of using the proposed model as part of a broader AI governance model, both society and academia can be considered in terms of society's role when answering an agency's inquiry regarding the ethical behaviour of AI systems.

Principles of robotics (Boden et al., 2017)

After pinpointing the responsibilities of all agents involved in robotics, five principles were established in a guideline for robot designers, manufacturers, and users. The main goal of the rules is to emphasise that robots are tools, whereas humans are the actual responsible agents. The proposed rules are:

a. Robots should not be designed as weapons, except in the interests of national security.
b. Robots should be designed and operated to comply with existing laws, including those dealing with privacy.
c. Robots should be designed to be safe and secure.
d. Robots should not be used to exploit vulnerable users by pretending to feel emotions.
e. It should be possible to find out who is responsible for any particular robot.

Aiming to encourage responsibility within the robot-related research and the industrial community, seven messages have been created to highlight the responsible innovation spirit needed to abide by the rules.

The opportunity to use this proposal in audits performed by regulatory agencies can be identified, and that need must be reflected in the legislation to be adapted or created.

Agile AI governance (Wallach & Marchant, 2018)

Aware of the concerns regarding AI impacts exceeding the regulatory scope, capabilities, and jurisdiction of an agency or nation, the authors propose a model to address this governance challenge.

The model predicts actions performed by a Governance Coordinating Committee at the national level and a Global Governance Coordinating Committee. The main goal is a soft-law strategy that mitigates risks while the legislation is being drawn up. The soft governance part involves industry standards, social codes, labs, certification practices, procedures, and programmes. The hard governance part concentrates on laws, regulations, and regulatory groups.

A national committee would coordinate the efforts of a governance process encompassing stakeholders to produce recommendations, reports, and roadmaps, while monitoring those actions at the same time. This national forum would also be a perfect structure to enforce soft governance mechanisms as a necessary complement to the hard ones.

On the international level, a global committee would not only coordinate agreements among countries, but also establish a common understanding of which international standards should be used as a soft governance strategy. The international approach is also advocated to bring some balance to the several countries that are not yet participating in the AI regulation dynamics, considering that the current situation makes them more vulnerable.

The proposed model takes a relationship network into account to address AI in a way that bolsters the formulation of actual standards while the legislation matures. The agile meaning of this governance is its incremental approach, which allows for continuous inputs. This would be an alternative to the problem posed by the temporal mismatch between formal regulatory actions and the production and commercialisation of deep machine learning-based products and services around the world. The success of this proposal depends on the amount of effort put into it by the market, academia, government, insurance companies, and organised civil society.

Sustainable AI development (Djeffal, 2018)

Considering the closer connection between sustainable development and governance, the author highlights that governance mechanisms are built to be continuously improved. The proposal concerns the entire lifecycle of an AI-based solution as the main foundation of a Sustainable AI Development (SAID) framework.

Analysed under the lens of a governance structure, SAID is stratified into the following layers: Technological, Social, and Governance.

At the base, the technology layer is in charge of specific applications involving architecture, data, and algorithm design.

Focusing on the impacts systems have on society, the social layer deals with the process of inserting technology into real life. It encompasses an analysis of the potential consequences of using AI in the social sphere.

Highlighting the importance of a broad treatment, the governance layer looks at the way algorithms influence both national and international decisions.

SAID gathers the different approaches examined in the various frameworks and somehow materialises the perception that, in order to be effective, AI regulation demands actions by IT and Social Sciences (Law, Business Administration, Philosophy, and Psychology) professionals alike. It also reminds us that, due to the topic's complexity, an AI governance model must include different process tiers.

Ethical framework for automation using robotics (Wright & Schultz, 2018)

Concerned with the integration between several stakeholders and automation using AI, this framework integrates the Stakeholders Theory with the Social Contract Theory in an attempt to find ethical grounds for developing, providing, and utilising AI.

The proposal considers as stakeholders: workers, the market, governments, the economy, and society in general. The impacts on the job market, from an ethical perspective, and the relationships among those stakeholders are highly emphasised.

The framework is based on a set of steps ranging from the identification of stakeholders, analysis of the social contracts among them, an assessment of how stakeholders are impacted, and lastly, actions aimed at mitigating the risk of terminating or breaching work contracts. An important target to be reached is increasing the benefits for stakeholders.

It is worth noting that this proposal considers as stakeholders those workers whose jobs or occupations will be modified with the introduction of AI into products and services. Due to the complexity of interests among stakeholders and all the labour concerns, the framework fits in the government policy-making process. The impact of such public policies on the country's economy may result in the need for laws, which means the legislative branch must be included as a stakeholder.

Intelligent model to regulate learning algorithms (Buiten, 2019)

Focused on a strategy to fight intelligent services that contain biases, this model postulates that an algorithm should assess the essential elements of a machine learning process (data, testing algorithms, and decision models). The proposal is founded on the thesis that the transparency of a code is insufficient to guarantee an unbiased solution and admits that it is still possible to find biases, even when learning from vast amounts of data.

In the data domain, all data samples are assumed to include some built-in biases that need to be considered. The data must be checked to ensure their validity, reliability, and proper data dependency.

Regarding the testing algorithms, the model recommends using a variety of algorithms and comparing their performance. However, that must be done only after discovering the quality of the available data.

The decision-making process is seen as a delicate phase in which developers must be aware of the correlations between variables, because hidden relationships may obscure a biased orientation. It also acknowledges the difficulty of identifying those problems automatically as algorithms grow in complexity.

Universal declaration of human rights as a framework (Donahoe & Metzger, 2019)

This model is founded on the argument that the several different frameworks related to each specific area of ethics are insufficient to regulate AI on an international scale, both in the private sector and within the government. Due to that gap, the Universal Declaration of Human Rights (Kunz, 1949) has been considered a mature approach that different cultures have been adopting for decades. Modern adjustments were made by the UN Human Rights Council in 2011, published as the UN Guiding Principles on Business and Human Rights (United Nations, 2011), which highlight the roles and responsibilities of private-sector businesses in the protection of human rights.

Under the human rights framework, governments have the duty to protect citizens from violations and infringements of their rights by other governments and non-State actors, including the private sector. Donahoe and Metzger's proposal deals with the centrality of the human person as the focal point of governance and society. It seeks to address the potential impacts of AI, such as:

- The right to equal protection and non-discrimination—avoiding biases in the data and ensuring fairness in machine-based decisions.
- The right to life and personal security—concerning autonomous weapons that move beyond human control.
- The right to an effective remedy for violations and infringements of rights—transparency, fairness, and accountability in cases where AI systems impact people's rights.
- The right to privacy—addressing the loss of privacy in data-driven societies and the need to protect personally identifiable data.
- The rights to work and to enjoy an adequate standard of living—guiding governance decisions around the displacement of human workers by AI.

Software requirement model for the ethical assessment of robots (Millar, 2016)

Considering ethics as a social enterprise, the proposal puts forth a set of general specifications to be considered in a system aimed at assessing robots during their construction. To that effect, five major rules have been built:

- Balancing designer and user requirements, considering the potential damages.
- Utilising a user-centred ethical evaluation tool for AI systems, which must use design methodologies that are able to identify the impacts on human values in use contexts.
- Including the psychology of user-robot relationship variables in the ethical evaluation tool to identify variables such as the user's emotional state.
- Compliance with the Human-Robotics Interaction Code of Ethics (Riek & Howard, 2014).
- Designers' understanding of both acceptable and unacceptable design features, which could be implemented by including ethicists in design teams.

It seems the proposal may be utilised by the industry and regulatory agencies alike. In both cases, it could be the first red flag signalling the need for a red button in robot projects (Arnold & Scheutz, 2018).

Ethical judgement model for codes (Bonnemains et al., 2018)

Considering that (a) an ethical framework allows us to deal with situations involving ethical dilemmas, (b) one framework alone is not efficient enough to compute an ethical decision, and (c) tackling ethical decisions is better than avoiding them, the author proposes a formal logical model that can be implemented by an agent facing an ethical dilemma, with the ability to both make decisions and explain those decisions. It assumes that formal expression analyses are especially useful to identify the subjectivity of a decision.

Different judgements on possible decisions have been studied according to three ethical frameworks: consequentialist ethics, deontological ethics, and the Doctrine of Double Effect. In the path toward a refined and final framework, various ethical dilemmas have been formalised in judgment functions that return three possible results: acceptable ($\top$), unacceptable ($\bot$), or undetermined (?). The concepts of 'decision', 'event', and 'effect' were taken into account when building the model's functionalities.

Those analyses can be appreciated when we judge someone or something based on particular moral theories.

Asilomar AI principles (Future of Life Institute, 2019b)

The governance model proposed by the Asilomar Conference resulted in 23 AI Principles undersigned by thousands of experts (Kozuka, 2019). Grouped under "Research Issues", "Ethics and Values", and "Longer-Term Issues", those principles encompass the lifecycle of an AI-embedded product or service—from motivation and funding to the assessment of benefits and judgement criteria concerning its impacts.

In the Research Issues dimension, the recommendations are to: research goals and funding, establish a connection between researchers and policymakers, research the culture of cooperation, and promote synergy to avoid corner-cutting when devising safety standards.

In the Ethics and Values dimensions, the orientations are to: maintain AI systems secure during their entire lifecycle, make them transparent in case of failure as well as in judgment results, consider designers and builders of advanced AI systems as stakeholders in the responsibility chain, align AI systems' values with their users', design AI systems to be compatible with human rights and cultural diversity, preserve personal privacy, share their benefits as much as possible, and make it possible for a human to take control of AI systems, if so desired.

Finally, in the Longer-Term Issues sphere, the principles are to: be cautious when making decisions without a consensus, build a mitigation plan to deal with the risks, plan a recursive type of self-improvement, and develop AI systems based only on widely shared ethical ideals.

European ethics guidelines for trustworthy AI (AI HLEG 2019b)

With the goal of creating guidelines to orient a new AI governance, a team of experts entitled High-Level Expert Group on Artificial Intelligence has drawn up the Ethics Guidelines for a Trustworthy AI for the European Commission based on a structure supported by values that should be considered throughout the system's lifecycle: lawful, ethical, and robust AI.

Based on the European Union Charter of Fundamental Rights (EU Parliament, 2012), the model establishes trustworthy AI as a key element for a governance framework (Kozuka, 2019) has been built using a three-tier structure.

The highest tier addresses ethical principles based on fundamental human rights: respect for human autonomy, prevention of damages, fairness, and explicability. To ensure fairness in a society with different interests and objectives, it defends an explicable decision-making process. It should consider traceability, auditability, and transparent communications regarding system capabilities. It also recommends that particular attention be paid to vulnerable groups and situations characterised by asymmetries of power or information (employers and workers, or businesses and consumers).

The second tier includes the key requirements necessary for implementing an AI-based system or service throughout its lifecycle: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; social and environmental wellbeing; and accountability. All requirements are connected to one another through a full-mesh relationship where each one of them has the same weight.

Special attention is suggested to the oversight as part of a governance mechanism that could use human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC). A strong connection can also be found between "privacy and data governance" and "diversity, non-discrimination, and fairness", due to the need for mechanisms to avoid inadvertent historical biases, incompleteness, and inadequate data governance models. Regarding the accountability concerns, a recommendation is given to carry out an impact assessment prior to and during the development.

Defending a trustworthy AI implementation throughout the lifecycle of an AI system, the model demands a process-oriented approach that encompasses both technical and non-technical methods when implementing the requirements. Within the non-technical approach, one can find legislation and corporate guidelines encompassing codes of conduct, policies, performance indicators, and agreed-upon standards. Those standards consider AI users, consumers, organisations, research institutions, and governments as stakeholders. They also include a certification granted to organisations that produce transparent, accountable, and fair AI systems in accordance with the established standards. The entity in charge of the certification could play an important role in the communications with "industry and/or public oversight groups, sharing best practices, discussing dilemmas, or reporting emerging issues of ethical concerns."

For the base tier, a list of recommendations directed at the operationalisation of the key requirements in the upper tier for each specific system has been formulated.

Ethically aligned design (IEEE 2019)

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems has proposed five general principles for AI systems and a guideline recommending actions to establish ethical and social implementations for intelligent and autonomous systems that prioritise human wellbeing.

According to that model, the ethical design, development, and implementation of AI systems should consider the following principles: human rights, wellbeing, accountability, transparency, and awareness of misuse.

Focusing on personal data rights, the wellbeing promoted by the effects on the economy, the legal frameworks for accountability and transparency, and the education and awareness policies, recommendations were made to a wide set of stakeholders.

To governments: the governance framework should include standards and regulatory agencies, provide society with ethics education and security awareness regarding the potential risks, improve digital literacy, use multiple metrics as wellbeing indicators, and implement a wellbeing impact assessment.

To industries: programmatic levels of accountability should be provided to address culpability in legal matters, transparency by design, intelligibility of a system's operation and decisions, damage mitigation strategies, assessment starting at the design phase, understanding how each jurisdiction would treat the damage caused by a given AI system.

To the legislative sphere: responsibility, culpability, liability, and accountability issues should be classified.

General recommendations: certification of AI systems, identification and prioritisation of standards for each category of AI systems, continuously updating the standards, metrics utilised to assess AI systems, agreements on moral decisions, evaluation by third parties, applying the classical methodologies of deontological and teleological ethics to machine learning, adherence to the code of conduct by the AI production team, and bridging the language gap between technologists, philosophers, and policymakers.

At the international level: establishing a global multi-stakeholder dialogue to determine the best practices, facilitating AI research and development in developing nations, and using indicators to assess AI-related technological interventions in those countries.

Government and industries: identifying the types of decisions and operations that should never be delegated to AI systems.

A special guideline for implementing an ethical culture in organisations (IEEE, 2020) has also been built, encompassing a strategy to assess the level of each dimension to be developed (lagging, basic, advanced, and leading).

Avoiding biases and discrimination (Lin et al., 2020)

In order to amplify the effectiveness of bias-reduction intervention procedures in cases of implicit biases, the framework explores an innovative AI-assisted intervention based on a bidimensional approach.

In the first dimension, the different types of information AI provide to users are captured: the current state of affairs (descriptive information), the likelihood of future states (predictive information), and the expected utility of an action (prescriptive information). It considers that all interventions are prescriptive, and the knowledge-based systems (KBS) will decide to intervene depending on how they simulate the results.

In the second dimension, an AI system can intervene in different phases of the decision-making process (input-based interventions, output-based interventions, and cognition-based interventions) as part of an interactive process.

It is a case of regulation by software, which could be used by the industry and service providers as part of their internal process.

Standardisation exchange model (Lewis et al., 2020)

Considering the importance of standardisation in a regulation strategy, the model proposes a process among functional entities in the AI value chain through which information related to standards is exchanged among them.

Classified by their functional roles, the actors—data providers, AI system creator, AI system operator, AI user, oversight authority, and associate stakeholder—change standards focusing on a trustworthy AI.

The benefits of each exchange are presented, as well as the potential topics for new standardisations. Most of them concern issues to be considered in an AI product certification process.

Although the focus is on the industry, the model considers the importance of the government in the whole process and the need for an international community to discuss the standards.

Algorithmic impact assessment (Canadian Government 2020)

Aiming to help public and private-sector companies assess and mitigate the impacts of deploying an automated decision-making system, the Canadian Government has developed the Algorithmic Impact Assessment (AIA) based on the Government Directive on Automated Decision-Making. The AIA questionnaire considers the reasons for using AI on decision-making processes, the capabilities encompassed by the system, algorithm transparency and explainability, system category (health, social assistance, economic, etc.), development and training process, system and data architecture, stakeholders, and risk mitigation measures.

The impact assessment addresses the four levels according to how the decisions impact the rights, health, or well-being, the economic interests of individuals or communities, and the ongoing sustainability of an ecosystem. Thus, levels I, II, III, and IV are each related to a certain impact, namely, reversible brief, reversible in the short term, difficult to reverse, and irreversible.

The Directive on Automated Decision-Making was designed by the Canadian Government to make its administrative decisions compatible with core administrative law principles, such as transparency, accountability, legality, and procedural fairness.

The requirements considered by the Directive on Automated Decision-Making are distributed between two pillars: transparency and quality assurance. Among the transparency requirements, it establishes that:

- Notice on relevant websites must be issued before decisions are made,
- Meaningful explanations must be provided to affected individuals regarding the decisions made,
- The Government of Canada has the right to access all components of the system.

Among the quality assurance requirements, there are rules to ensure testing and monitoring outcomes, data quality, peer review, employee training, contingency, security, compliance with the law, and human intervention.

AI governance by human rights-centred design, deliberation and oversight (Yeung et al., 2019)

Considering international human rights-based standards as the most promising governance framework to deal with ethical standards, Yeung et al. (2019) have proposed the Human Rights-Centred Design, Deliberation, and Oversight model to deal with AI-related ethical issues with legal support. Based on a global approach, the proposed model integrates a suit of technical, organisational, and evaluation tools and techniques involving many stakeholders.

The proposal presents norms based on human rights as the foundation for ethical standards with which AI systems must demonstrably comply:

a. Design and development that take stakeholders' opinions into account. In case an assessment has resulted in "high" or "very high" risks to human rights, a redesign should be pursued.
b. Formal assessment and testing to evaluate their compliance with human rights-based standards. It would occur regularly during the entire life cycle of a system's development—design, specification, prototyping, development, and implementation. A systematic and periodic post-implementation monitoring would be established, through which the AI system would be submitted for review by sending out the related documentation and reports to a public authority.
c. Independent oversight by an external, technically competent entity invested with legal investigation and sanction powers.
d. Auditability supported by traceability and by evidence that the AI system is operating as desired and that it was properly documented during its entire life cycle of development.

The authors highlight the need for laws and norms encompassing all steps covered by the model.

Good AI society (AI4People, 2018)

Focused on the establishment of a good AI society, the proposal joins ethical principles and specific recommendations to enable stakeholders to seize opportunities and avoid or minimise risks.

The model encompasses five ethical principles: Beneficence, Non-maleficence, Autonomy, Justice, and Explicability.

The recommendations are categorised as: assessment, development, incentivisation, and support.

- Assessing institutions on their capacity to reduce the mistakes made by AI systems.
- Considering existing legislation, using participatory mechanisms to align with social values, and assessing tasks/decision-making that should not be delegated to AI systems.
- Assessing current regulations to provide a legislative framework that could keep pace with technological developments.
- Developing a framework to enhance the explicability of AI systems.
- Developing legal procedures to permit the scrutiny of algorithmic decisions in court.
- Developing auditing mechanisms for AI systems to identify unwanted consequences.
- Developing a process to remedy or compensate for damage caused by AI.
- Developing agreed-upon metrics for the trustworthiness of AI products and services.
- Developing a new EU oversight agency responsible for the scientific evaluation and supervision of AI products and services.
- Developing a European observatory for AI.
- Developing legal instruments to prepare and adjust the work environment to the changes brought about by AI.
- Financially incentivising a socially preferable development and use of AI.
- Financially incentivising cross-disciplinary cooperation in the fields of technology, social issues, legal studies, and ethics.
- Incentivising a regular review of the legislation to foster socially positive innovation.
- Financially incentivising the use of lawfully special zones for empirical testing and development.
- Financially incentivising research on the public perception of AI.
- Supporting self-regulatory codes of conduct for data- and AI-related professionals.
- Supporting corporate boards of directors to take responsibility for the ethical implications of AI technologies in their organisations.

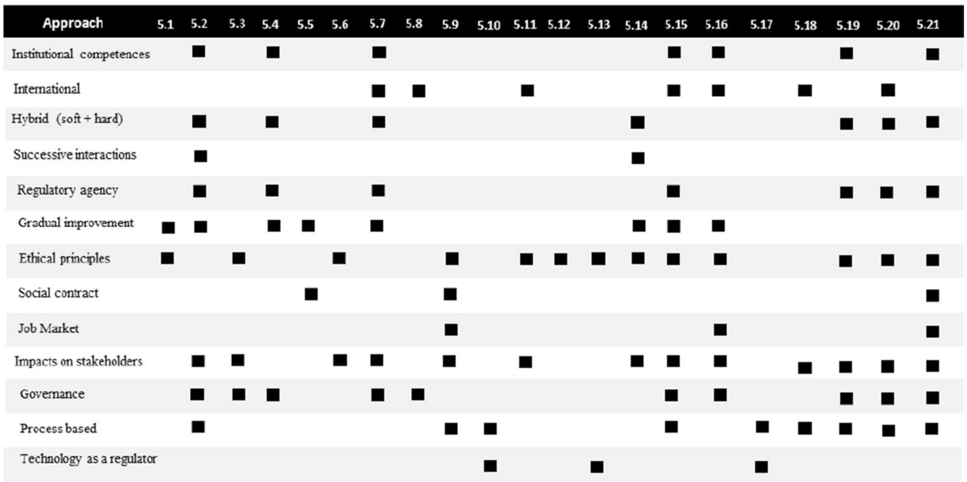
Framework approaches

An analysis of the approaches adopted by each of the 21 frameworks proposed in the sample resulted in Table 1.

The fact that ethical guidelines exist is not enough to have any effect on the software development industry. Thus, models that are strongly grounded on ethical principles require legal mechanisms to fulfill those recommendations (Hagendorff, 2019).

Frameworks that encompass the competencies of government institutions have also foreseen the existence of a regulatory agency, as well as the need for mechanisms to

Table 1 Comparative table of the approaches explored in the frameworks, compiled by author.



Approach	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9	5.10	5.11	5.12	5.13	5.14	5.15	5.16	5.17	5.18	5.19	5.20	5.21
Institutional competences		■		■			■								■	■			■		■
International							■	■			■				■	■		■		■	
Hybrid (soft + hard)		■		■			■							■					■	■	■
Successive interactions		■												■							
Regulatory agency		■		■			■								■				■	■	■
Gradual improvement	■	■		■	■		■							■	■	■					
Ethical principles	■		■			■			■		■	■	■	■	■	■			■	■	■
Social contract					■				■												■
Job Market									■							■					■
Impacts on stakeholders		■	■			■	■		■		■			■	■	■		■	■	■	■
Governance		■	■	■			■	■							■	■			■	■	■
Process based		■							■	■					■		■	■	■	■	■
Technology as a regulator										■			■				■				

help the legislative branch speed up its law-making process, aiming for a safer and faster AI regulation.

Frameworks that take the social contract into account rank among the most open to society's participation in a co-production with the government. Those models consider citizens as outstanding stakeholders. Concerns over the impacts on the job market are also a way to assess the impact on stakeholders.

The main argument that proposes a gradual deployment of the regulation is a risk mitigation strategy, but it could also be combined with successive interactions between the legislative branch and the regulatory agency, thus enabling continuous improvement during the legislative procedure.

The interactive regulatory governance model, the agile governance, the ethics guideline for trustworthy AI, the ethically aligned design, the algorithmic impact assessment, the good AI society, and the AI governance by human rights-centred design, deliberation, and oversight proposals encompass a larger number of topics. The AI HLEG proposal highlights that a trustworthy AI must be lawful, ethical, and robust. The others explore the relationship among all parties involved in the regulation process and the attempt to find balance between more or less rigid or flexible mechanisms. It is worth noting that the agile governance proposal does not exclude conventional actions for a formal regulation—the interactive regulatory governance model and the competency-based regulatory model, both of which involve the legislative branch. Therefore, this configures a transitional situation in which consensual standards would be agreed upon and enforced, and the risks would be mitigated until legal mechanisms are made official, which is very similar to the concept of Dynamic Regulation, in which feedback serves as a basis for the maturity of the regulatory instrument (Kaal & Vermeulen, 2017).

When analysing several movements advocating the establishment of criteria for best using AI, studies identified an opportunity to develop a competition around a technological reform (Greene et al., 2019). Pondering over the need to find synergy among global AI regulation-oriented actions, a few proposals rely on a worldwide effort, which sometimes is described as an international committee, while other times just as a joint effort by governments and multinational companies.

Despite the small number of existing software-based regulation models, similar models are likely to arise, since the increasing complexity of AI solutions results in more system rules (Lamo & Calo, 2018; Liu, 2017; Prakken, 2017; Verheij, 2016), which in turn means a higher likelihood of conflicts among those rules in combined systems (Bench-Capon & Modgil, 2017).

AI regulatory and governance framework

The supplementary nature of some of the models confirms the perception that the impacts of AI would demand a combination of design, laws, and education (Calo, 2011). When debating over the complexity of a framework to address such a multidisciplinary topic (Bonnemais et al., 2018) embedded into the political and social context (Leitner & Stiefmueller, 2019), an AI regulatory and governance framework—AIR—was built to include the main contributions from each model in the examined sample (Fig. 1).

Focused on reducing the gap between ethical principles and actions by each stakeholder and making the relationship among them in different dimensions of knowledge clearer, the AIR framework is based on a wide governance process.

Although the government's exclusive competencies are highlighted, aiming for more accuracy in its actions, the power of the State has been distributed among the legislative, the executive, and the judicial branches. This

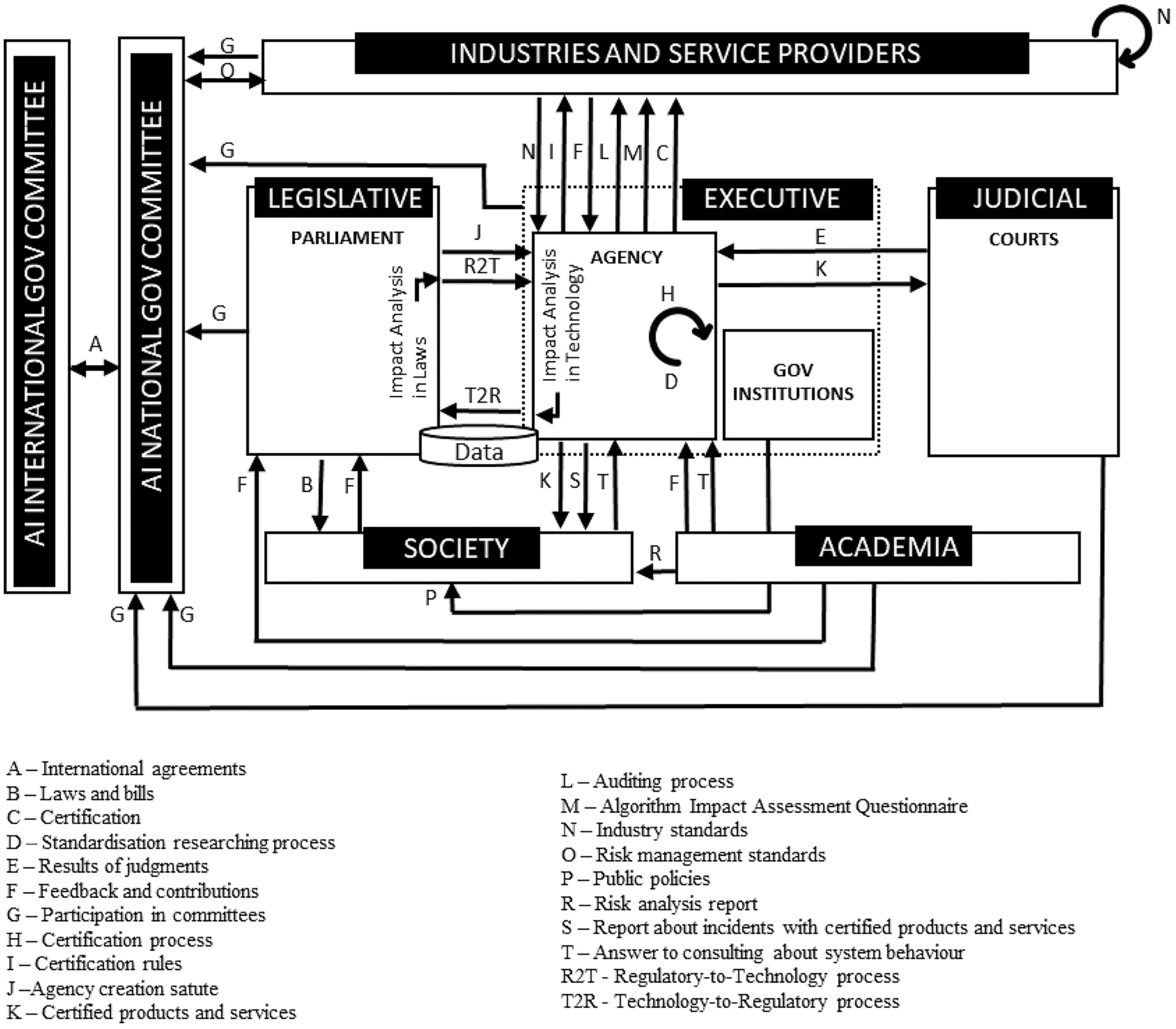

INDUSTRIES AND SERVICE PROVIDERS
N
G
O
AI INTERNATIONAL GOV COMMITTEE
AI NATIONAL GOV COMMITTEE
A
G
G
LEGISLATIVE
PARLIAMENT
Impact Analysis in Laws
J
R2T
T2R
Data
N
I
F
L
M
C
EXECUTIVE
AGENCY
Impact Analysis in Technology
H
D
GOV INSTITUTIONS
JUDICIAL
COURTS
E
K
F
B
F
K
S
T
F
T
SOCIETY
R
ACADEMIA
P
G
G
A – International agreements
B – Laws and bills
C – Certification
D – Standardisation researching process
E – Results of judgments
F – Feedback and contributions
G – Participation in committees
H – Certification process
I – Certification rules
J –Agency creation satute
K – Certified products and services
L – Auditing process
M – Algorithm Impact Assessment Questionnaire
N – Industry standards
O – Risk management standards
P – Public policies
R – Risk analysis report
S – Report about incidents with certified products and services
T – Answer to consulting about system behaviour
R2T - Regulatory-to-Technology process
T2R - Technology-to-Regulatory process

Fig. 1 AIR framework

segmentation is used by many countries as a functional way to distribute power, according to which the legislative creates the laws, the executive enforces those laws, and the judicial is in charge of solving whatever conflicts arise to guarantee justice and law abidance (Maluf, 1995).

Apart from making laws, it is crucial to maintain the legislative branch open so that its bills (B) can be discussed with society, receiving constant feedback and contributions not only through e-participation systems, but also through a special channel established with scholars, who could also attend the legislative committee meetings (F).

The Parliament or Congress, as an instance of the legislative branch, would approve a statute (J) to create an AI regulatory agency as part of the Federal Government (executive branch). This could be a good moment to define AI, or at least to demand that the agency do it.

Upon its creation, the regulatory agency would establish a strong relationship with the Parliament as part of an ongoing process in which the legislative would survey the impact on the legislation and its evolution based on the knowledge obtained from the regulatory agency (T2R—Technology-To-Regulatory), much like the regulatory agency structures its internal work processes based on the legislation discussed and approved by the legislative (R2T—Regulatory-To-Technology).

The T2R is necessary, at least until each new category of AI systems has been deeply studied by the regulatory agency. Due to the complexity and specificity of AI services and products, laws could potentially be created for each specific field. The natural evolution of the former would also cause the latter to evolve in the long term. A practical way to implement the T2R flow is through the regulatory agency frequently attending the legislative committee meetings to discuss AI regulation.

As a complement to T2R, the R2T flow would be started at least when a new version of a bill is discussed at the legislative committee meetings and when a new law is approved. R2T also feeds other internal processes of the regulatory agency in order to update them with the legislative understanding of what can be regulated by law, which can trigger three reactions: (a) alerts regarding the limitations that the bill/law brings to the ongoing projects of the industries and service providers; (b) opportunities to expand the standards by discussing them with the industries and service providers; and (c) updating certification and auditing processes with new compliance issues.

Among the regulatory agency's competencies, a couple of processes require speed and synergy: analysing how the legislation affects the technology process and standardising the research, certification, and auditing processes. In order to be effective, those processes must consider a huge number of variables involved in the entire lifecycle of an AI system: design, prototyping, development, testing, deployment, commercialisation, and use. The efficiency and knowledge of the regulatory agency are expected to possess depend on mechanisms that support those processes (Fig. 2): formal representation models for ethical dilemmas, impact on stakeholders' assessment according to ethical principles, data governance assessment, development process assessment, systems to identify biased machine learning, and assessment of risk mitigation measures.

Being responsible for a closer interaction between the Parliament and the regulatory agency due to the R2T flow, the analyses of how the legislation affects the technology process must be corroborated with information structures that are able to represent the law based on a technical mindset. Those involved must be skilled in both areas of knowledge. The quality and efficiency of this synchronicity of mixed mindsets are strengthened by means of a data repository shared by the Parliament and the regulatory agency. Examples of such data would include: issues related to whether AI projects/products comply with the law, AI project/product impact assessment, legislation/regulation collected overtime across AI projects/products, ethical committee decisions upon approval requests, AI project/product and regulatory assessments under both private and public guidelines.

In order to offer controlled autonomy to the AI industry, technical standards must be established while bills are being discussed. An agile interaction between the "industries and service providers" and the regulatory agency is supported by the standardisation of the research process (D). Despite being a process inside the regulatory agency, standardising research actions entails an in-depth study of ethical and safe mechanisms to make the new projects seen in the AI market feasible. A very technically skilled staff must be allocated to that task, which requires robust laboratories.

As a strategy to motivate the AI market to follow the best practices, standards, and laws (when they exist), the regulatory agency would certify products and services using a certification process (H). Companies that submit their products to the regulatory agency, after a successful appraisal, would receive a certificate (C) within their field of action (transport, healthcare, entertainment, education, military,

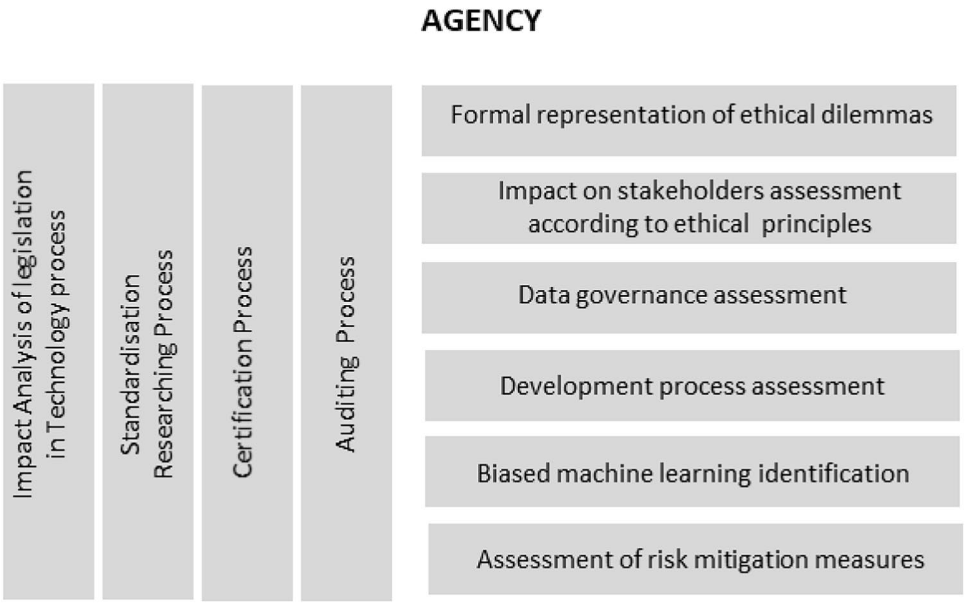

AGENCY
Impact Analysis of legislation in Technology process
Standardisation Researching Process
Certification Process
Auditing Process
Formal representation of ethical dilemmas
Impact on stakeholders assessment according to ethical principles
Data governance assessment
Development process assessment
Biased machine learning identification
Assessment of risk mitigation measures

Fig. 2 Regulatory agency in the AIR framework

etc.). The strictness and nature of the assessment process could be different for each of those fields. It is also a way to communicate to society where people can place their trust when buying or using an AI system.

Through a quick process, the industries and service providers would need to receive the regulatory agency's certification rules stated as clearly as possible (I), while providing feedback (F) on the conditions that preclude the development process required by the regulatory agency from moving forward. Accountability requirements would be assured if those lists would show not only new products and services that have been certified, but also those that have lost their certification.

The issuance of certificates could be a strategy to be applied before laws are passed, since they already inform society, in a transparent fashion, about the safety levels and risks of the products and services it consumes. Advertising campaigns by the government and certified companies would also strengthen that strategy.

A robust strategy to avoid fake certifications would be desirable, such as a blockchain mechanism implemented by the agency containing the updated certification list for a given country. Aiming to increase citizen trust in AI certification, certificates could be issued using a digital signature in the system's code, setting an attribute associated with that specific code version. In case someone wants to know if the version of a commercialised AI system is updated, all they need to do is compare the digital signature with the one that is available on the regulatory agency's website.

The regulatory agency could make an "algorithm impact assessment questionnaire" (M) available to the industries and government institutions in order to offer a simulation tool through which they could know, in advance, their level of compliance. It would also fit as a preparation stage for a certification submission.

And finally, an auditing process (L) would be supported by the regulatory agency to check companies demanding certification and certified companies that need to update their certification, as well as to verify issues demanded by courts in case of sentences related to damages supposedly caused by AI systems. This audit would take place in five dimensions: impact on stakeholders based on ethical principles, data governance, development process models, identification of biased machine learning, and risk mitigation measures. The auditing process should be part of regular monitoring through which not only internal changes in companies, but also future problems coming from new arrangements in society could be identified.

Any failures or damages noticed in a certified AI product or service must trigger an internal audit process to identify whether there were problems or limitations in other agency processes that could be a reminder of internal improvement. In a broader, more transparent fashion, the agency should publish the audit results and the next steps (S).

The regulatory agency's processes are interconnected through a knowledge stemming from the mechanisms shown in Fig. 2, which should be handled as much as possible by a skilled multidisciplinary team, since the ethical and technological dimensions are mixed.

Mechanisms for formally representing ethical dilemmas are important to create a transparent communication channel between ethicists and technical profiles. It could also help distinguish between the part of the decision-making algorithm that is related to a dilemma and the rest of the code in relation to which there is a consensus regarding the best decision. This representation model is expected to be continually improving as society changes and new dilemmas are identified. This analysis is interconnected with the impact on the stakeholders' assessment according to ethical principles.

As each company has its own system development process, the regulatory agency must have a process to guarantee a broad system development process assessment, probably by attempting to measure the sample against the best practices and the risks related to each step that does not follow them.

Since a biased machine learning can result from problems with data collection, testing algorithms, or decision models, the regulatory agency must consider all those phases in its development process assessment models. A data governance assessment is an important analysis that is connected to the system development process as well as to the biased machine learning identification process.

The results of the regulatory agency's analysis materialise the total sum of all risks identified in an evaluated AI product or service for which there should be a risk mitigation plan.

As the regulatory agency is a natural actor to create and communicate the best practices to the industries and service providers, the agency must be aware of all projects and trends in the AI market, otherwise companies will not adopt those practices. An alternative to mitigate that risk is to strengthen the dialogue with the industries and service providers on the purpose of contributing to industry standards (N), thus allowing technology to improve its development while the legislation is still under debate, or in case it is not necessary. On the industries' and service providers' side, in order to increase the probability of a successful investment, a gradual strategy supported by a governance model should be behind the implementation of those good practices. Industry standards (N) must incorporate all parameters that are needed for communication among the "industries and service providers" along the entire value chain of an AI system.

As happens in Parliament, an open practice by the regulatory agency is likewise desirable, receiving feedback from academia (F). That feedback and those contributions, among other information, could be how society perceives ethical behaviours. A partnership between academia and the regulatory agency, combining scholars and researchers in the agency's staff, could be a sustainable alternative for maintaining a highly skilled team of professionals dealing with many processes simultaneously.

At an advanced level of an AI governance model, a "society-on-the loop" mechanism could be structured to collect the evaluation of a certain category of AI systems based on their behaviour using an ethical approach. Both civil society and academia could accomplish this. The answers (T) would feed the regulatory agency in the form of a survey to identify potential opportunities for improvement in its internal processes.

Regardless of the existence of a "society-on-the loop" mechanism, academia is always a good, reliable source of risk analysis reports (R) to be published periodically.

Law enforcement by courts would also undergo a continuous learning process with regard to interpretations based on the legislation in effect, as well as on new laws. In countries where the certification is incorporated into laws, decisions on cases involving uncertified companies would be treated differently from those involving certified companies. Thus, society and the courts would need to have up-to-date information about each company's certified products and services (K). Considering a continuous learning process, the regulatory agency would receive the judgment decisions of all cases involving AI systems (E), which would then be stored in the data repository shared with the Parliament. The decisions on cases may indicate types of AI technology use that the regulatory agency has not researched yet, and they may also indicate the need for changes in the legislation. A significant challenge would be to identify when an incident is avoidable or not. In those situations, experts must be involved in the investigation to find out the purpose of supporting the courts.

In order to balance the equation that rules the job market on the path to a digital economy, the government may create public policies (P) to make it feasible to implement in a timely manner the changes required in employer and student skills. Public policies might also be necessary to maintain an advertising campaign to inform people about the importance of certification and standards for AI products and services, helping them to identify when there is a potential case of an AI-embedded system.

As usual, public policies are a long-term strategy that may require actions by different government institutions, but there are many alternatives for implementing them, depending on the country. The regulatory agency may also provide government institutions with information about where and how those changes are needed. In some cases, by means of the T2R flow, the agency may notify the Parliament that a law is lacking that better regulates public policies.

On a national level, discussions to facilitate priority actions and the recognition of industry standards would be enabled through an AI Governance Committee, bringing together the public and private sectors (G). The synergy of efforts for the benefit of all stakeholders must be established, since many variables are considered. Beyond the regulatory agency, other government institutions would probably participate in this national committee, due to the wide impact its decisions could have. For instance, building human capacity and preparing the labour market transformation is a decision that might require a strategy that impacts many ministries and state governments. Adjustments to the current legislation related to many different subjects should probably be made to support the whole transition.

We should not forget the committee's governance approach, which requires working with indicators, i.e. data produced by its stakeholders. Therefore, a national AI governance committee would require at least collection, storage, and analysis processes within other institutions and businesses.

The agreed-upon standards (N) make it possible to move forward in some technological dimensions, while the Parliament discusses adjustments to the legislation when necessary. The risk management criteria (O) related to the use of those standards would be negotiated between the national committee and the industries and service providers, since each standard could impact a long productive chain.

The plethora of components in AI services and products of global reach imposes actions that would be agreed upon in an International Governance Committee comprising representatives from each country's committee (A). On many occasions, transparency in production processes is only feasible through complex international agreements, because corporate trade policies must adapt to different countries. A global strategy could be established to facilitate the production and delivery of standards, as well as the dissemination of best practices in undeveloped countries, since without that help the gap between them and the countries in which an AI governance has been established would increase hugely, putting them in a fragile position. In that regard, one should keep in mind that international standards are not limited to technological issues. Further, those standards also incorporate ethical principles, despite any cultural differences. The Universal Declaration of Human Rights could be a global base to engage governments to face the challenge of dealing with differences among national legislations.

The expert skills and engagement power of self-regulated organisations are a rich contribution to the international AI governance committee.

A possible adjustment entails the segmentation of tasks in charge of the regulatory agency, sharing them with or transferring them to other government institutions. For instance, the audit process could also be implemented by different government institutions in charge of auditing cases of discrimination using personal data, or investigations related to the development of autonomous weapons in that country. Hence, it is important to highlight that laws such as the EU GDPR (2016) only affect personal data. Nonetheless, AI discrimination risks have a wider reach than personal data.

Sharing the standardisation process with specialised private-sector organisations could also be an alternative. In that case, the connection between the standardisation process and the other regulatory agency processes should be maintained.

Despite being represented as a unique institution, the regulatory agency could be materialised as a group of agencies distributed across the country. To that effect, partnerships among countries could also allow for the creation of a set of agencies sharing resources, processes, and knowledge. In both cases, agencies could specialise in different categories of AI products and services. Although the certification issued for a specific category is independent of the certificate issued for another category, a communication process among the agencies is needed to increase the knowledge of how each AI product/service behaves and evolves over time.

Another adjustment to how the AIR framework is interpreted relates to what can be classified as “industries and service providers”. Private-sector companies are considered first. However, since any organisation that develops AI systems or offers services based on AI systems would fall in that category, public organisations may also be included.

Conclusion

The need and urgency to regulate Artificial Intelligence seem indisputable. The complexity of the topic is also evident, whether due to the advanced nature of technology or because its impacts structurally affect social standards. This combination materialises the perception of a problem that is yet to be completely defined.

A study of the literature through a sample comprising 109 documents (articles, laws, and government strategies) revealed significant efforts to identify and scale the risks and ethical dilemmas related to AI, as well as to seek a model for regulating AI based on different methodologies.

The heterogeneous nature of the professional profiles involved in the debate evinces the complexity and maturity with which the topic is being studied. Such an in-depth approach, on the one hand, may have caused certain delays in research, but on the other, it has prevented inappropriate regulatory solutions from being made official.

We had also seen the birth of a reshaped perception of the legislation, as had occurred with disruptive innovations in the past, when legislative efforts focused on adapting laws to the new paradigms brought about by electricity, telephone, and computers. Since this is a more difficult challenge, AI lawmakers will consider that we are still starting to discover the applications of smart algorithms. Therefore, a balance must be kept between a rigid damage prevention and technological development strategy (Gurkaynak et al., 2016).

Despite all efforts being directed to AI regulation and governance, there is still an expressive gap between ethical principles and a functional model that is able to encompass all areas of knowledge that are necessary to deal with the required complexity. The 21 proposed models found in the sample are based on supplementary approaches and are therefore insufficient when analysed separately. Due to the heterogeneous nature of those skills and interests, an ideal model should harmonise interests, offering benefits to all stakeholders during the entire lifecycle of an AI product or service.

The consolidation and process orientation approach proposed by the AIR framework (Fig. 1) seems to be the most adequate strategy for the deployment of an AI governance, given the existence of several agents and the laterality of the topic, which intertwines different areas of knowledge. The expanded view of the presented AIR framework will enable all agents involved to identify their role in the governance process, while establishing a roadmap for a gradual and uninterrupted deployment.

It also contributes to the creation of a new reward and punishment model to balance out this new reality (Bryson, 2018; Waser, 2015), taking into account the world as it will be (Lin et al., 2011).

On the path to improve each component of the AIR framework, more than bringing them closer together, there needs to be a synchronisation of stakeholders towards a sustainable regulation. Along that journey, an alliance between scholars and the government’s three agents (the executive, legislative, and judicial branches) is crucial for the macroprocess of regulation.

The countries leading the debate are probably ready to coordinate the partnerships and agreements among institutions that are necessary for a comprehensive and effective governance, as well as to initiate a regulation process. Nonetheless, the launch of AI-embedded products in countries that have advanced regulation models, in and of itself, does not guarantee the same safety levels for countries that are still unripe in this regard.

Much is yet to happen in the formulation of solutions using real-case scenarios to enable an empirical analysis and studies of the evolution of the models presented in the examined sample. To that effect, the AIR framework can make it tangible and feasible to synchronise all the stakeholders’ efforts to achieve an effective result, thus culminating in the creation of a reference model of AI governance in which maturity levels would be established that could be monitored by international bodies in a collaborative action. The way we and future generations will live our lives depends on that cooperation.

References

Aayog, N. (2018). National Strategy for Artificial Intelligence: #AI for All (Discussion Paper) https://www.niti.gov.in/writeaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf. Accessed 30 July 2020.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access Review, 6*, 52138–52160

AI HLEG - High-Level Expert Group on Artificial Intelligence. (2019a). A definition of AI: Main capabilities and disciplines. Definition developed for the purpose of the AI HLEG's deliverables.

AI HLEG - High-Level Expert Group on Artificial. (2019b). Ethics guidelines for trustworthy AI. High-level expert group on artificial intelligence for the European Commission.

AI4People. (2018). Ethical framework for a good society: opportunities, risks, principles, and recommendations. *Atomium – European Institute for Science, Media and Democracy.* http://www.eismd.eu/wp-content/uploads/2019/02/Ethical-Framework-for-a-Good-AI-Society.pdf. Accessed 21 June 2019.

Amigoni, F., & Schiaffonati, V. (2018). Ethics for robots as experimental technologies. *IEEE Robotics & Automation Magazine, 25*, 30–36

Arkin, R. C. (2011). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. In *Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction*, 121–128.

Arnold, T., & Scheutz, M. (2018). The big red button is too late: an alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology, 20*, 59–69

Beltran, N. (2020). Artificial intelligence in Lethal Autonomous Weapon Systems: What's the problem? Uppsala University – Department of Theology.

Bench-Capon, T., & Modgil, S. (2017). Norms and value based reasoning: Justifying compliance and violation. *Artificial Intelligence & Law Review, 25*, 29–64

Benjamins, V. R. & García I. S. (2020). Towards a framework for understanding societal and ethical implications of Artificial Intelligence. *Vulnerabilidad y cultura digital* by Dykinson. pp 87–98.

Black, J. (2002) Critical reflections on regulation. *Australian Journal of Legal Philosopy, 27*, 1–35. http://www.austlii.edu.au/au/journals/AUJlLegPhil/2002/1.pdf. Accessed 30 July 2020.

Bonnemais, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: Some technical and ethical challenges. *Ethics and Information Technology, 20*, 41–58

Buiten, C. M. (2019). Towards intelligent regulation of artificial intelligence. *European Journal of Risk Regulation, 10*(1), 41–59

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorrell, T., Wallis, M., Whitby, B., & Winfield, A. (2017). Principles of robotics: Regulating robots in the real world. *Connection Science, 29*(2), 124–129

Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: Some technical and ethical challenges. *Ethics Information Technology., 20*, 41

Borgesius, F. Z. (2018). Discrimination, artificial intelligence, and algorithmic decision-making. Study for the Council of Europe.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bryson, J. J. (2018). Patiency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology, 20*, 15–26

Butterworth, M. (2018). The ICO and artificial intelligence: The role of fairness in the GDPR framework. *Computer Law & Security Review, 34*, 257–268

Calo, M. R. (2011). Peeping hals. *Artificial Intelligence Review, 175*, 940–994

Calo, M. R. (2015). Robotics and the lessons of cyberlaw. *California Law Review, 103*(3), 513–563

Caron, M. S., & Gupta, A. (2020). The social contract for AI. Cornell University. https://arxiv.org/abs/2006.08140v1 Accessed 6 Dec 2020.

Canada Government. (2020). Algorithmic Impact Assessment. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html. Accessed 15 Dec 2020.

Carter, D. (2020). Regulation and ethics in artificial intelligence and machine learning technologies: Where are we now? Who is responsible? Can the information professional play a role? *Business Information Review., 37*(2), 60–68

Cath, C., Watcher, S., Mittelsadt, B., Taddeo, M., & Floridi, L. (2017). Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. https://ssrn.com/abstract=2906249 or https://doi.org/10.2139/ssrn.2906249. Accessed 21 June 2019.

Cave, S., Nyrup, R., Vold, K., & Weller, A. (2019). Motivations and risks of machine ethics. *Proceedings of the IEEE, 107*(3), 562–574

Cerka, P., Grigiene, J., & Sirbikite, G. (2015). Liability for damages caused by artificial intelligence. *Computer Law & Security Review, 31*(3), 376–389

Cerka, P., Grigiene, J., & Sirbikyte, G. (2017). Is it possible to grant legal personality to artificial intelligence software systems? *Computer Law & Security Review, 33*(5), 685–699

Conitzer, V., Sinnott-Armstrong, W., Borg, J. S, Deng, Y., & Kramer, M. (2017). Moral decision making for artificial intelligence. *AAAI Publication, 31º Conference on Artificial Intelligence*

Council of Europe. (2018). European commission for the efficiency of justice, 'European ethical charter on the use of artificial intelligence in judicial systems and their environment. https://rm.coe.int/ethical-charter-en-for-publication-4-december-2018/16808f699c. Accessed 30 July 2020.

Davis, E. (2015). Ethical guidelines for a superintelligence. *Artificial Intelligence Review, 220*, 121–124

Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems, 77*, 1–14

Djeffal, C. (2018). Sustainable AI Development (SAID): On the road to more access to justice. https://ssrn.com/abstract=3298980 or https://doi.org/10.2139/ssrn.3298980. Accessed 30 July 2020.

Donahoe, E., & Metzger, M. M. (2019). Artificial intelligence and human rights. *Journal of Democracy, 30*(2), 115–126

Dubai (2019). Smart Dubai. Artificial intelligence principles and ethics. https://smartdubai.ae/initiatives/ai-principles-ethics. Accessed 20 July 2020.

EU GDPR. (2016). European Parliament. General Data Protection Regulation. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679. Accessed 30 July 2020.

EU Parliament. (2012). Charter of Fundamental Rights of the European Union (2012/C 326/02), *Official Journal of the European Union*, 2012 C 326, (pp. 391).

European Commission. (2019). Communication from the Commission to the European Parliament, the Council, The European Economic and Social Committee and the Committee of the Regions. Brussels. https://www.eea.europa.eu/policy-documents/communication-from-the-commission-to-1. Accessed 30 July 2020.

Firth-Butterfield, K. (2017). Artificial Intelligence and the Law: More questions than answers. *Scitech Lawyer, 14*, 28–31

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A. C., & Srikumar, M. (2020). Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI. Berkman Klein Center for Internet & Society.

Floridi, L., Cowls, J., King, T., & Taddeo, M. (2020). How to design AI for social good: seven essential factors. *Science and Engineering Ethics, 26*, 1771

French, P. M. (2018). For a Meaningful Artificial Intelligence: Toward a French and European Strategy. Mission assigned by the French Prime Minister. https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf. Accessed 30 July 2020.

Future of Life Institute. (2019a). National and International AI Strategies. https://futureoflife.org/national-international-ai-strategies/. Accessed 20 September 2019.

Future of Life Institute. (2019b). Ansilomar AI Principles. https://futureoflife.org/ai-principles/. Accessed 20 September 2019.

German Federal Government. (2018). German Federal Ministry of Education and Research, the Federal Ministry for Economic Affairs and Energy, and the Federal Ministry of Labor and Social Affairs. Artificial Intelligence Strategy. https://www.ki-strategie-deutschland.de/home.html. Accessed 30 July 2020.

Greene, D., Hoffmann, A. L., & Stark, L. (2019). Better, nicer, clearer, fairer: A critical assessment of the movement for ethical artificial intelligence and machine learning. *Hawaii International Conference on System Sciences* 52nd, 2019.

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly, 30*(3), 611–642

Gurkaynak, G., Yilmaz, I., & Haksever, G. (2016). Stifling artificial intelligence: Human perils. *Computer Law & Security Review, 32*(5), 749–758

Hagendorff, T. (2019). The ethics of AI ethics: An evaluation of guidelines. CoRR, abs/1903.03425.

Hilb, M. (2020). Toward artificial governance? The role of artificial intelligence in shaping the future or corporate governance. *Journal of Management and Governance.*

Hildebrandt, M. (2018). Algorithmic regulation and the rule of law. *Philosophy Transactions of the Royal Society,* 376 (2128).

Holder, C., Khurana, V., Harrison, F., & Jacobs, L. (2016a). Robotics and law: Key legal and regulatory implications of the robotics age (Part I of II). *Computer Law & Security Review, 32*(3), 383–402

Holder, C., Khurana, V., Hook, J., Bacon, G., & Day, R. (2016b). Robotics and law: key legal and regulatory implications of the robotics age (Part II of II). *Computer Law Secure Review, 32*, 557–576

House of Lords. (2018). AI in the UK: Ready, willing and able? *Select Committee on Artificial Intelligence*, Report of Session 2017–19. 13 March 2018.

IEEE. (2019). Ethically Aligned Design. Committees of The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2nd version. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf. Accessed 20 July 2020

IEEE. (2020). a call to action for business using AI—Ethically aligned design for business. https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead/ead-for-business.pdf. Accessed 20 July 2020.

Jackson, B. W. (2019). Artificial Intelligence and the Fog of Innovation: A deep-dive on governance and the liability of autonomous systems. 35 *Santa Clara High Tech*. L.J. 35

Jackson, B. W. (2020). Cybersecurity, privacy, and artificial intelligence: An examination of legal issues surrounding the European Union General Data Protection Regulation and Autonomous Network Defense, 21 *Minnesota Journal of Law, Science & Technology*, *21*

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399. https://doi.org/10.1038/s42256-019-0088-2Accessed20July2020

Japanese Cabinet Office. (2019). Social principles of human-centric artificial intelligence. Council for science, technology and innovation, https://www8.cao.go.jp/cstp/english/humancentricai.pdf. Accessed 20 July 2020

Kaal, W. A., & Vermeulen, E. P.M. (2017). How to regulate disruptive innovation: From facts to data. *Jurimetrics*, *57*(2).

Kozuka, S. (2019). A governance framework for the development and use of artificial intelligence: Lessons from the comparison of Japanese and European initiatives. *Uniform Law Review, 24*, 315–329

Kunz, J. (1949). The United Nations declaration of human rights. *American Journal of International Law, 43*(2), 316–323

Larsson, S. (2020). On the governance of artificial intelligence through ethics guidelines. *Asian Journal of Law and Society*, 1–23.

Lenardon, J. P. A. (2017). The Regulation of Artificial Intelligence. *Master Thesis. Tilburg Institute for Law, Technology and Society*. Netherlands.

Lewis, D., Hogan, L., Filip, D., & Wall, P. J. (2020). Global challenges in the standardization of ethics for trustworthy AI. https://doi.org/10.5281/zenodo.3516525. Accessed 30 July 2020.

Lamo, M. & Calo, R. (2018). Regulating Bot Speech. *UCLA Law Review 2019*, July 16, 2018.

Leitner, C., & Stiefmueller, C. M. (2019). Disruptive technologies and the public sector: The changing dynamics of governance. In A. Baimenov & P. Liverakos (Eds.), *Public service excellence in the 21st century.* (pp. 238–239). Palgrave Macmillan.

Lewis, T., & Yildirim, H. (2002). Learning by doing and dynamic regulation. *The RAND Journal of Economics*, *33*(1), 22–36. www.jstor.org/stable/2696373 Accessed 20 July 2020.

Lin, P., Abney, K., & Bekey, G. (2011). Robot ethics: Mapping the issues for a mechanized world. *Artificial Intelligence Review, 175*, 942–949

Lin, Y., Hung, T., & Huang, L. T. (2020). Engineering equity: How AI can help reduce the harm of implicit bias. *Philosophy & Technology*. https://doi.org/10.1007/s13347-020-00406-7

Liu, H. (2017). Irresponsibilities, inequalities and injustice for autonomous vehicles. *Ethics Information Technology Journal, 19*, 193–207

Maluf S. (1995). *Teoria Geral do Estado*. 23ª ed., 205–208. Editora Saraiva. São Paulo.

Mantelero, A. (2018). AI & Big Data: A blueprint for human rights, social and ethical impact assessment. *Computer Law & Security Review, 34*(4), 754–772

Mika, N., Nadezhda, G., Jaana, L., & Raija, K., (2019). Ethical AI for the governance of the Society: Challenges and opportunities. *CEUR Workshop Proceedings*, *2505*, 20–26. http://ceur-ws.org/Vol-2505/paper03.pdf. Accessed 20 July 2020.

Millar, J. (2016). An ethics evaluation tool for automating ethical decision-making in robots and self-driving cars. *Applied Artificial Intelligence, 30*(8), 787–809

Monetary Authority of Singapore. (2019). Monetary Authority of Singapore. Principles to promote fairness, ethics, accountability and transparency (FEAT) in the use of artificial intelligence and data analytics in Singapore's Financial Sector. https://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/

Monographs%20and%20Information%20Papers/FEAT%20 Principles%20Final.pdf. Accessed 20 July 2020

Nevejans, N. (2016). European civil law rules in robotics. Study requested by the European Parliament's Committee on Legal Affairs. *Policy Department Citizens' Right and Constitutional Affairs*.

Neznamov, A. V. (2020). Regulatory landscape of artificial intelligence advances in social science, education and humanities research, *volume* 420 pp 201–204. XVII *International Research-to-Practice Conference 2020*. Atlantatis Press.

Organisation for Economic Co-operation and Development (2019). 'Recommendation of the Council on Artificial Intelligence'.

Partnership on AI to Benefit People and Society. (2016) https://www.partnershiponai.org/about/. Accessed 12 July 2020.

Pedro, A. P. (2014). Ética, moral, axiologia e valores: confusões e ambiguidades em torno de um conceito comum. *Kriterion*, vol. 55. Belo Horizonte, nº 130, Dez./2014, 483–498.

Poel, I. V. (2016). An ethical framework for evaluating experimental technology. *Science and Engineering Ethics, 22*(3), 667–686

Prakken, H. (2017). On the problem of making autonomous vehicles conform to traffic law. *Artificial Intelligence & Law, 25*, 341–363

Rahwan, I. (2017). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology., 20*, 5–14

Reed, C. (2018). How should we regulate artificial intelligence? *Philosophy Transactions of the Royal Society, 376*, 2128

Riek, L. D., & Howard, D. (2014). A code of ethics for human-robot interaction profession proceedings of we robot, 2014. SSRN: https://ssrn.com/abstract=2757805. Accessed 20 July 2020.

Rousseau, J. (2016). *The Social Contract*. (202–230). ISBN: 978911495741. London: Sovereign.

Russell, S., & Norvig, P. (1995). *Artificial Intelligence. A Modern Approach*. (pp. 4–5). Prentice Hall.

Scherer, M. U. (2016). Regulating artificial intelligence systems: Risks, challenges, competences and strategies. *Harvard Journal of Law & Technology, 29*(2), 354–398

Schrader, D., & Ghosh, D. (2018). Proactively protecting against the singularity: Ethical decision making AI. *IEEE Computer and Reliability Societies Review, 16*(3), 56–63

Smuha, N. A. (2020). *Beyond a human rights-based approach to AI governance: Promise*. Philosophy & Technology.

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good: An ethical framework will help to harness the potential of AI while keeping humans in control. *Science Review, 361*(6404), 751–752

Toronto. (2020). The Toronto declaration: Protecting the right to equality and non-discrimination in machine learning systems. https://www.torontodeclaration.org/. Accessed 20 July 2020

Tutt, A. (2017). An FDA for algorithms. *Administrative Law Review, 69*(83), 83–123

UK Government. (2018). Government response to House of Lords Artificial Intelligence Select Committee's Report on AI in the UK: Ready, Willing and Able? https://www.gov.uk/government/publications/ai-in-the-uk-ready-willing-and-able-government-response-to-the-select-committee-report. Accessed 31 December 2020

United Nations. (2011). *UN guiding principles on business and human rights*. (p. 2011). UN Human Rights Council.

University of Montreal. (2018). Montreal Declaration for a Responsible Development of Artificial Intelligence. https://www.montrealdeclaration-responsibleai.com/the-declaration Accessed 20 July 2020

US Congress. (2019). H.Res.153 - Supporting the de7velopment of guidelines for ethical development of artificial intelligence. https://www.congress.gov/bill/116th-congress/house-resolution/153?q=%7B%22search%22%3A%5B%22ARTIFICIAL+INTELLIGENCE%22%5D%7D&s=2&r=4

US Congress. (2020). s.3891 – Advancing Artificial Intelligence Research Act of 2020. https://www.congress.gov/bill/116th-congress/senate-bill/3891?q=%7B%22search%22%3A%5B%22ARTIFICIAL+INTELLIGENCE%22%5D%7D&s=3&r=7

Villaronga, E. F., & Heldeweg, M. (2018). Regulation, I presume? Said the robot: Towards an iterative regulatory process for robot governance. *Computer Law & Security Review*, 21 June, 2018.

Verheij, B. (2016). Formalizing value-guided argumentation for ethical systems design. *Artificial Intelligence & Law Review, 24*(4), 387–407

Yeung, K., Howes, A., & Pogrebna, G. (2019). AI governance by human rights-centred design, deliberation and oversight: An end to ethics washing (June 21, 2019). Forthcoming in M Dubber and F Pasquale (eds.) *The Oxford Handbook of AI Ethics*, Oxford University Press (2019), https://doi.org/10.2139/ssrn.3435011. Accessed 15 December 2020.

Wallach, W., & Marchant, G. E. (2018). An agile ethical/legal model for the international and national governance of ai and robotics. *Association for the Advancement of Artificial Intelligence*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6191666/. Accessed 20 July 2020

Waser, M. (2015). Designing, implementing and enforcing a coherent system of laws, ethics and morals for intelligent machines (including humans). *Procedia Computer Science, 71*, 106–111

Wright, S. A., & Schultz, A. (2018). The rising tide of artificial intelligence and business automation: Developing an ethical framework. *Business Horizons, 61*(6), 823–832

Journal of the Knowledge Economy (2024) 15:20547–20577
https://doi.org/10.1007/s13132-024-01974-1

The Over-Concentration of Innovation and Firm-Specific Knowledge in the Artificial Intelligence Industry

Pedro Jácome de Moura Jr.[1] · Carlos Denner dos Santos Junior[2] · Carlo Gabriel Porto-Bellini[1] · José Jorge Lima Dias Junior[1]

Received: 6 February 2023 / Accepted: 4 April 2024 / Published online: 16 April 2024

Abstract

The development of the artificial intelligence (AI) landscape has been impressive in virtually all economic sectors in recent years. Our study discusses the over-concentration of AI knowledge (OCAIK) as the origin of dominance over the global AI industry by a small number of companies and universities that deploy the needed resources to develop and use cutting edge, inimitable AI knowledge. Business agents appropriate AI-related scholarly research and absorb research findings that grant them increasingly inimitable competitive advantages over new entrants. Our study verifies the occurrence of OCAIK by processing thousands of papers presented in AI conferences from 2013 to 2022. To analyze our hypotheses, we used classification techniques and inferential statistics. We found a significant difference between clusters of companies that we called ordinary investors and outlier investors. We also observed the influence of universities in the correlation between OCAIK and investments made in both research and development (R&D) and capital goods. Our findings indicate a strong collaboration between AI leading companies and universities in generating firm-specific AI knowledge. We additionally offer novel insights on the resource-based view (RBV) and the knowledge-based view (KBV) research traditions, in that business competition may reach a point of no return if only incremental innovation is devised instead of radical innovation to break the chains of knowledge accumulation and technological implementation by a strict number of agents.

Keywords Artificial intelligence · Competitive advantage · Research and development · Knowledge-based view · Oligopolies

This article is part of the Topical Collection on *AI in the Knowledge Economy and Society: Implications for Theory, Policy and Practice*

Extended author information available on the last page of the article

Introduction

The world has approximately 214 million companies in various fields of industry, agriculture, trade, and services (Statista, 2022). They compete by means of differentiation to gain relative advantage over competitors and, consequently, market share. One of the main strategic actions to obtain competitive advantage has been to invest in research and development (R&D) (Boiko, 2021; Menke, 1997; Nayak et al., 2022). About a trillion dollars is invested annually in R&D by the 2500 companies that lead R&D in the world, corresponding to 90% of the world's business-funded R&D (Grassano et al., 2022). Knowledge has been long considered resourceful to leverage the needed differentiation and, ultimately, competitiveness (Barney, 1991; Grant, 1996). In particular, firm-specific knowledge is better for business advantage than publicly available knowledge, since the proprietary nature of firm-specific knowledge makes it more difficult to absorb, imitate, or replace (Wang et al., 2016). Technology thus emerges as key to acquire and exploit firm-specific knowledge (Nayak et al., 2022), particularly the technology that allows for big data analytics (Dahiya et al., 2021) and artificial intelligence (AI) (Fredström et al., 2021).

Towards competitive advantage, companies may want to establish partnership with universities as universities have the necessary background to conduct research in emerging, knowledge-based fields such as AI (Shao et al., 2020). Moreover, given the focus of universities on teaching, basic research, and community outreach, i.e., they assume a more non-commercial mission, universities do not represent serious threats to businesses (Choi & Contractor, 2019). Also, to exploit business advantages, universities face greater difficulties to transform resources and skills into real capabilities or core competences (Mahdi et al., 2019). The latter may be due to the fact that while business companies focus on the integration of resources (e.g., research findings) and the exploitation of advantages acquired by the combination of resources, universities in their turn focus on the production of factual knowledge (to promote an open debate within the academic community as well as to advance the frontiers of scholarly knowledge). Moreover, universities are not necessarily interested in the application of integrated resources. To illustrate this point, take Microsoft's *Optics for the Cloud* project, which aims to revolutionize data storage in partnership with multiple US and UK universities (Parmigiani et al., 2021), where each university carries out specific studies and whose results are integrated with results from other universities under Microsoft's discretion and power. This is an example that universities work in the periphery of business initiatives. The peripheral role in the business landscape is a consequence of companies seeking profit and competitive advantage, whereas universities focus on research rather than on marketing a product (D'Este & Perkmann, 2010). Another reason for the peripheral role of universities is that even when universities focus their entrepreneurial activities on technology licensing through spin-off firms, the spin-offs are often bought soon after by large companies as a way to gain access to cutting-edge knowledge (Ferreira et al., 2018).

AI, in particular, has received attention from governments and scholars regarding ethical issues, regulation, and governance (e.g., Kerr et al., 2020; Roberts et al., 2021). An example of governmental concern regarding AI ethics, regulation, and governance is the *AI Watch Report* launched by the European Commission in December 2018 to monitor the impact of AI as a source of potential social implications (Benetta et al., 2021). On the academic side, a related initiative is the *AI Governance: A Research Agenda Report*, published by the Future of Humanity Institute (FHI), University of Oxford, which highlights risks of a potential oligopolistic global market structure motivated by interests on AI (Dafoe, 2018). Such an oligopolistic structure manifests when only a few companies can invest what is needed regarding capital and R&D to keep pace with constant technical iterations (Ding & Dafoe, 2021). Such issues lead us to assume that companies' accumulated R&D investment capacity, when reinforced by capabilities generated through partnerships with universities, determines the level of firm-specific AI knowledge and competitive performance. The relationship between investments and competitive performance is in fact well documented in the information technology (IT) literature. For example, IT as a knowledge-intensive industry uses its capabilities on combining IT resources to create competitive advantages (Bharadwaj, 2000). Since resources cannot offer competitive advantages by themselves (Cohen & Olsen, 2013), knowledge, skills, and partnerships are examples of complementary capabilities that explain a firm's superior performance (Fink, 2011) and could possibly explain the current advances in AI. In fact, in the light of the knowledge-based view (KBV), knowledge, skills, and partnerships are needed for firm-specific knowledge integration into organizational capability (Grant, 1996), which makes KBV a promising theory for analyzing the *over-concentration of AI knowledge* (hereafter referred to as OCAIK).

While studies on AI investments have focused mostly on regional statistics (e.g., Jeon et al., 2024; Benetta et al., 2021; and Dafoe, 2018), the role of large companies in shaping the AI landscape has been anecdotal (e.g., Newman, 2017; Webb, 2019), leading us to elaborate the following research questions:

RQ1 To what extent do companies that invest heavily in firm-specific AI knowledge differentiate from those that do not?
RQ2 How do universities collaborate with companies to reinforce OCAIK?

By answering these questions, we expect to contribute in two ways for scholarly and applied knowledge. First, by discussing OCAIK, the study sheds light on the need to revisit the role of independently or publicly funded research institutions, so that those institutions become cognizant of OCAIK and champion policies for the regulation and democratization of access to AI resources. And second, the study adds to the KBV literature regarding how much advantage is needed for sustainable, competitive advantage.

We developed a parsimonious model to answer RQ1 and RQ2. Specifically, we test the positive relationship between (1) R&D investments by companies and the formation of a group of AI leading companies (LC); (2) R&D expenditure by universities and the formation of a group of AI leading universities (LU); (3) the partnership between LC and LU, and the production of LC firm-specific AI knowledge; and (4) the partnership between LC and universities in general, and the production of LC firm-specific AI knowledge.

The article is organized as follows. First, we discuss the theoretical framing of competitive advantage, the KBV, and the role of AI developments in that framing, especially machine learning. Second, we present a methodological approach that uses secondary data on R&D investments and academic production (papers published in leading AI conferences) to answer the research questions. Third, we discuss the results along with implications for theory and practice, limitations, and future studies. And fourth, we present conclusions about OCAIK and its potential consequences.

Literature Review

Competitive Advantage and the Knowledge-Based View of the Firm (KBV)

Competitive advantage is "a multifaceted construct arising out of diverse contextual manifestations" mostly regarding an economic perspective and grounded on rivalry and leadership (Nayak et al., 2022, p. 977). The concept has just recently gained a socio-technical perspective, as governments and industry become increasingly aware of societal expectations (Marakova et al., 2021) and of the role of emerging technologies for the competitive advantage at firm level (Shao et al., 2020) and country level (Ding & Dafoe, 2021). The resource-based view of the firm (RBV) comes to explain a firm's survival through its capacity of achieving differentiation in a competitive environment (Barney, 1991). In the light of the RBV, competitive advantage is achieved through cost reduction (e.g., scale gains, exclusive sources of raw materials, and a cost-effective workforce) or because of specific attributes of the firm's products or processes (e.g., trademarks, patents, network service, or distribution channels) (Newbert, 2008). The RBV assumes that (1) firms are composed of heterogeneous resources (the resources are not evenly distributed among all players); (2) firms will have superior performance if their resources are valuable, rare, difficult to imitate, irreplaceable, or difficult to move; and (3) the source of competitive advantage exists inside the firm. In their search for differentiation, firms develop organizational capabilities, i.e., a group of activities based on the development, flow, and exchange of information carried out systematically and allowing the firm to take advantage of its resources to generate valuable outcomes (Degravel, 2011). As such, the RBV expects certain stability in a competitive sector to give firms time to change and adapt (Teece et al., 1997).

However, in rapidly changing and unpredictable situations, knowledge emerges as the main source of competitive advantage by promoting the transformation of organizational capabilities into dynamic capabilities, i.e., "the organizational and strategic routines by which firms achieve new resource configurations as markets emerge, collide, split, evolve, and die" (Eisenhardt & Martin, 2000, p. 1107). The knowledge-based view of the firm (KBV) therefore extends the RBV by assuming that the need for dynamic capabilities is inherent to organizations and that knowledge is a superior resource, one which combines/recombines resources, including those that are external to the organization (Bharadwaj, 2000; Grant, 1996). The assumption of knowledge as a relevant organizational asset implies a strategic decision and further actions to promote curiosity to learn, adaptation/improvement of what has been learned, and exploitation to obtain benefits—what is known as *absorptive capacity* (Henard & McFadyen, 2006). The KBV thus explains competitive advantage by means of absorptive capacity, i.e., knowledge integration into organizational and dynamic capabilities (Zahra et al., 2020).

The KBV has been used to analyze R&D investments and competitive advantage in areas such as corporate social responsibility and sustainability (Ullah & Arslan, 2022), cybersecurity (Mongeau & Hajdasinski, 2021), emerging economies and national competitiveness (Ge & Liu, 2022), and in innovation in the healthcare sector (Orlando et al., 2021). However, it is not any sort of knowledge that promotes competitive advantage. Firm-specific knowledge has more potential for differentiation among competitors than publicly accessible knowledge (Barney, 1991; Grant, 1996), since the proprietary nature of firm-specific knowledge makes it more difficult to absorb, imitate, or replace (Pereira & Bamel, 2021; Wang et al., 2016). Firm-specific knowledge is composed of (1) expertise and skills of a firm's employees, manifesting through "common language, relationships, or a sense of identification that exists among departments within a firm," and increasing on the basis of employee involvement with prior related projects (Mayer et al., 2012, p. 3); (2) the levels of financial and human-resource slack for R&D (Wang et al., 2016); and (3) external sources and partnerships (Cai et al., 2019). Slack resources are also desirable for firm-specific knowledge development, since they exceed the minimum necessary for organizational operation, and, as such, work as "a buffer resource" in times of change (Yiu et al., 2020, p. 1210).

Companies that invest heavily in firm-specific knowledge are now combining their traditional innovation process with AI technologies to gain competitive advantage and differentiation vis-à-vis those who invest less (Bai & Li, 2020), specifically by applying AI to the development of firm's knowledge with potential for innovation in business models, product and service new features, innovation structure, market performance, and innovation in supply chain management (Bahoo et al., 2023).

The University-Industry-Government Helix

Academic findings may generate unique knowledge to be transferred to society through scholarly publications, but also by means of patenting, spin-off startups, joint research, and consulting (D'Este & Perkmann, 2010). Such modes of knowledge transfer have been long encouraged through public policies and government incentives (Etzkowitz & Leydesdorff, 1995). In fact, social agents urge universities to participate in collaborative innovation processes and knowledge transfer for social and economic development (Johnston, 2019). The *triple helix* metaphor illustrates the interaction that is expected to occur between universities, industry, and government (Etzkowitz & Leydesdorff, 1995). It is interesting for companies to collaborate with universities as universities focus on basic research rather than on profit, and as they do not pose threats to competition (Miotti & Sachwald, 2003). Universities are also of value for governments as universities promote open innovation and have the capacity to create and transfer knowledge to both the business and the public sectors (Johnston, 2019). Universities have thus been considered ideal partners for integration with governments and with the private sector in a tri-lateral collaborative arrangement, particularly in the domain of AI (Mikhaylov et al., 2018).

The traditional role of universities in teaching and research has been changing as a consequence of the competitive pressure universities face in the global economy (Ferreira et al., 2018), the decreasing availability of public funding (Miotto et al., 2020), and the idea that governments could stimulate universities to become entrepreneurial agents (Etzkowitz & Leydesdorff, 1995). Universities would then incorporate entrepreneurship in their institutional routines to obtain economic returns from knowledge creation (Formica, 2002). However, there is a lack of understanding on how knowledge is transferred and how collaborative partnerships are established, i.e., how companies select universities to partner with, and vice versa, as well as what are the underlying reasons for the partnership (Johnston, 2019).

Artificial Intelligence

AI emerged in the late 1940s (Bruderer, 2016) in seminal propositions on how to determine whether a digital computer could think like a human (Turing, 1950) and on how a computer could play human games, "something of the nature of judgement, and considerable trial and error, rather than a strict, unalterable computing process" (Shannon, 1950, p. 256). Shortly after, those propositions began to appear in practice, as described by Samuel (1959) when he wrote a computer program "to behave in a way which, if done by human beings or animals, would be described as involving the process of learning," that is, "[p]rogramming computers to learn from experience" (Samuel, 1959, p. 211). Still, in the early days, the *Eliza Effect* (Weizenbaum, 1966) introduced to the world further anthropomorphic characteristics of AI, particularly one that emulated human conversation.

Broadly, AI depends on the ability of machines to learn from experience, examples, and planned training, or, as more recently defined, "[a] computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E" (Mitchell, 1997, p. 2). The concept of AI encompasses (1) inputs (raw data, parameters), models (rules, heuristics, accuracy), and outcomes (predictions, decisions, trustworthiness) (Bui et al., 2020; Kaur et al., 2020); (2) technology and resources (hardware, software, data quality, security) (Challa et al., 2020; Hu et al., 2021); and (3) the influence on the context/environment (social, ethical, economic impacts) (Kerr et al., 2020; OECD, 2020). Such a wide scope of AI may be a consequence of the diversity of its development and applications. For instance, physicists and AI researchers work together to analyze and interpret "unmanageable volumes of data" produced by particle colliders (like the Large Hadron Collider) using specific models and algorithms to discover complex properties of matter (like the Higgs boson) (Castelvecchi, 2015, p. 18), and in medicine, the collaboration between physicians and AI researchers has led to the deployment of robots to assist remote surgery (Hamet & Tremblay, 2017) and nanorobots to precisely deliver drugs inside the human body (Hassanzadeh et al., 2019).

Machine learning (ML) has been the main tool for leveraging the complexity and usefulness of AI (Collins et al., 2021; Liu et al., 2021). ML evolved from simple pattern recognition through biological neuronal emulation (Rosenblatt, 1960; Uhr & Vossler, 1961) to recognition, identification, analysis, and classification of complex—and sometimes incomplete—contents, such as ancient texts (Assael et al., 2022). The profound developments in ML—nowadays under the umbrella term *deep learning*—have been carried out in countries with immense capacity to invest in research and infrastructure (Liu et al., 2021) with the support of research centers and universities (Shao et al., 2020).

However, while studies on AI investments have addressed the role of nations and governments, the role of companies has been anecdotal (e.g., Kovacevich, 2022; Newman, 2017; Webb, 2019). The state of the art in AI seems to have been defined by companies like Alphabet (Google), which develops TensorFlow, a powerful ecosystem for ML in use in many industries (Pang et al., 2020), and Microsoft/Open AI's chat-GPT, a tool for processing natural language that has become the standard in applications from search queries to language translation (Edwards, 2021). As another example, Meta delivered its SEER self-supervised computer vision model that can learn without data curation and labeling (Ramanathan et al., 2021), both of which are well-known pre-processing tasks in conventional computer vision training. In such a business environment, some authors identify a "race to AI" (Smuha, 2021, p. 3) leading to a "winner-takes-all" phenomenon (Ding & Dafoe, 2021, p. 192).

Hypotheses

Dominance over an industry favors the over-concentration of power either through the accumulation of knowledge (Marimon & Quadrini, 2011) or due to ownership of the most efficient resources (Arnosti & Weinberg, 2022), thus imposing barriers to competition. By the same token, we can expect that cutting-edge AI development is made possible by companies that possess the most efficient hardware as well as data in greater volume and variety. Assuming that investments in R&D and in equipment serve as a proxy for AI investments (Yiu et al., 2020; Benetta et al., 2021), we hypothesize:

> H_1: R&D investments by companies lead to the formation of a group of AI leading companies (LC).

Due to complexity and research costs, companies turn to partnerships in R&D (Choi & Contractor, 2019), preferably with universities (Mahdi et al., 2019). Although many universities participate in R&D with the industry, we can expect that some universities stand out due to the amount of resources they have available to invest in R&D, specifically R&D focused on computer science as a proxy for AI expenditures. Therefore, we hypothesize:

> H_2: R&D expenditure by universities leads to the formation of a group of AI leading universities (LU).

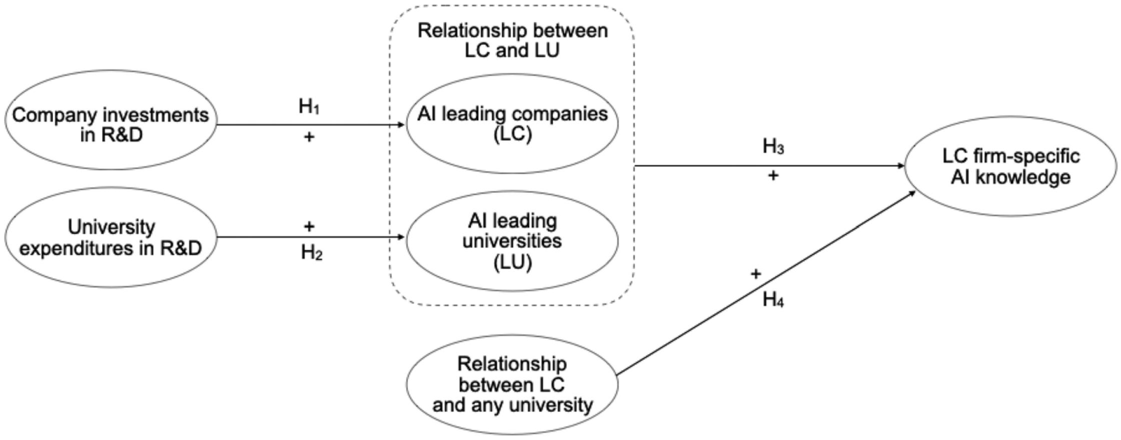

Relationship between LC and LU
Company investments in R&D
H1
+
AI leading companies (LC)
University expenditures in R&D
+
H2
AI leading universities (LU)
H3
+
LC firm-specific AI knowledge
+
H4
Relationship between LC and any university

Fig. 1 Conceptual model

Companies that occupy a central position in an interfirm research cooperation achieve greater efficacy in knowledge acquisition and innovation, and "[they] will likely benefit more from diverse partners" (Wang et al., 2020, p. 171), including universities and research centers. Such a strategic positioning is better spotted by the industry than by universities (Shao et al., 2020). Therefore, we expect that universities focus on the production of factual knowledge, but they do not necessarily consider the integration of resources, whereas companies devise the big picture of an organizational cooperation and the intended outcomes. Therefore, we hypothesize:

H_3: The partnership between LC and LU leads to the production of LC firm-specific AI knowledge.

Finally, universities are not uniform in producing knowledge (Huggins et al., 2012), with some universities being more prone than others to engaging in collaborative partnership with business companies (Johnston, 2019). This makes us expect that not only LU collaborate with LC for the production of AI knowledge, but other universities also do it. Therefore, we hypothesize:

H_4: The partnership between LC and universities in general leads to the production of LC firm-specific AI knowledge.

Figure 1 synthesizes the idea that institutions investing the most in R&D are the ones producing the most in AI, and this is done through collaborative strategies (partnerships). The underlying assumption is that a few companies produce relevant AI with the support of universities, which in turn gives rise to an over-concentration of firm-specific AI knowledge and barriers to competition.

Method

We used three sources of secondary data for the test of hypotheses. The first source contains R&D investments from companies (dataset "A"), the second source contains R&D expenditures from universities (dataset "B"), and the third source contains the academic output (papers presented in AI conferences) of companies when they produce alone as well as the academic output of companies in partnership with universities (dataset "C").

R&D investments have been adopted as a predictor for the operational capability in high-tech firms (e.g., Yiu et al., 2020), and a predictor for AI investments in the European context (e.g., Benetta et al., 2021). We followed the same approach for datasets A and B: dataset A comprises the European Commission's industrial R&D investment scoreboard (EC-IRI) for the 2500 companies that invested the largest sums in R&D worldwide each year; and dataset B comprises the National Science Foundation's Higher Education Research and Development (NSF-HERD) report on R&D expenditures from US universities, considering that US universities are among the best in all global performance rankings (Shanghai, THE, etc.). Both datasets A and B were accessed on March 2022, and they included data from 2012 to 2021 published each subsequent year by the EC-IRI and NSF-HERD reports.

The academic output (papers published in conference proceedings) has been adopted as a predictor for the relationship between companies and universities (e.g., Li et al., 2021). We followed this same approach to build dataset C from papers presented at conferences listed in the *2021-2022 International Conferences in Artificial Intelligence, Machine Learning, Computer Vision, Data Mining, Natural Language Processing and Robotics*[1] and provided that their h-index was at least 100. Six conference websites were accessed on August 2022, comprising data from 2013 to 2022: *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), *Neural Information Processing Systems* (NeurIPS), *International Conference on Computer Vision* (ICCV), *Annual Meeting of the Association for Computational Linguistics* (ACL), *Association for the Advancement of Artificial Intelligence* (AAAI), and *Conference on Empirical Methods in Natural Language Processing* (EMNLP).

As for the analysis of data, we adopted cluster analysis for H_1 and H_2, Web scraping and classification algorithms for H_3 and H_4, and the Mann-Whitney *U* for the test of hypotheses. Cluster analysis is a ML technique that does not require labeled data (non-supervised training) (Lai et al., 2019), which is the case for datasets A and B, and it offers a way to find patterns in raw data by grouping lines/observations with common characteristics (Li et al., 2020). Cluster analysis contributes to this study because it is suitable for grouping companies according to how closely associated their R&D investment capabilities are, which is required to answer the first research question.

Among the multiple classes of clustering algorithms available to test H1 and H2 (e.g., hierarchical, partition, grid, and density-based), we have opted for the density-based one, as it does not assume normality, does not depend on the shape of data, and deals well with noise (Lai et al., 2019). The density-based spatial clustering of applications with noise (DBSCAN; Martin et al., 2001) implements a density-based clustering and has been considered a standard in its class (Pelka, 2018; Fredström et al., 2021). DBSCAN does not require the *a priori* definition of the number of clusters to be found (Li et al., 2020), such as

[1] https://jackietseng.github.io/conference_call_for_paper/conferences-with-ccf.html

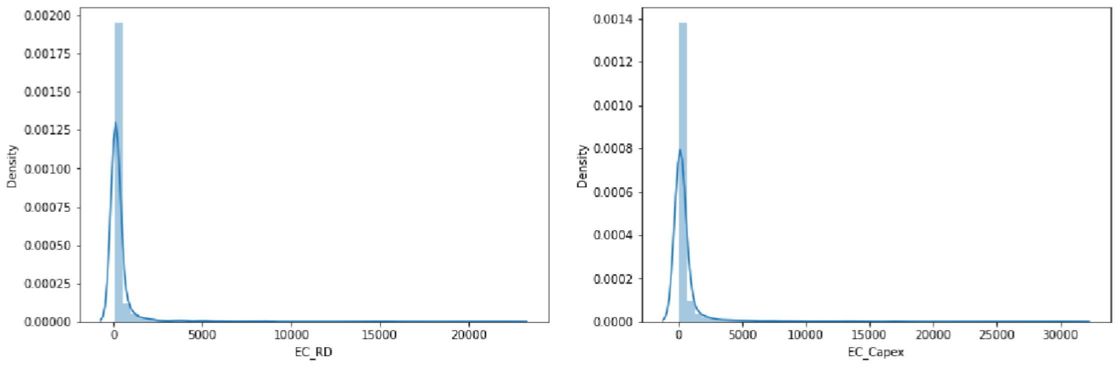


Fig. 2 Kurtosis analysis on R&D investments and Capex Source: developed by the authors

in the case of *k*-means. Instead, it requires two other parameters: Eps (the radius of a centroid) and MinPts (minimum amount of data points reachable within the centroid radius). Estimating both central parameters for the DBSCAN algorithm has been a challenge (Braune et al., 2015), to the point that many researchers try to automate such an estimation (e.g., Hou et al., 2016; Karami & Johansson, 2014), while others claim that this is a decision that is often based on a researcher's experience (Lai et al., 2019) through an iterative approach (Soni & Ganatra, 2016), sensitivity analysis (Fredström et al., 2021), or domain knowledge and heuristics (Schubert et al., 2017). Our approach followed Braune et al. (2015) and Soni and Ganatra (2016), assuming the outlier threshold as a determinant for both parameters.

The initial analysis of data suggested that the biggest investors are outliers (Fig. 2), which is a first important finding. Considering the presence of companies with massive investments in R&D and Capex (capital goods) in the data, we take each of those companies as a single cluster. Following such a rationale, MinPts=1 (at least one company/cluster) and Eps=2.5 (the outlier threshold radius—Euclidean distance—identified through iterative attempts). After finding the clusters, we applied Mann-Whitney tests to verify the statistical significance of differences as a non-parametric alternative to two independent sample *t*-tests.

For H_3 and H_4, we used Web scraping and specific classification algorithms. Web scraping is a technique to extract data directly from the World Wide Web using automated algorithms to convert the content of websites—usually, non-structured content—into structured datasets (Zhao, 2017). And the specific classification algorithms, in their turn, contribute to this study because we did not find any current technique to access and sort the conference papers and subsequently extract the precise information we needed to answer the second research question. We went through every Web page of AI conference proceedings in search of links to PDF files containing the papers. Then, we extracted the paper titles, authors, affiliations, and funding information (from the acknowledgments) to build dataset C. "Appendix" summarizes the procedures we have adopted for mining, collecting, and processing the three datasets, and it also explains how those sources were used to answer the research question.

Table 1 Top 20 largest R&D and Capex investors (2012–2021)

Company	Year									
	2021	**2020**	**2019**	**2018**	**2017**	**2016**	**2015**	**2014**	**2013**	**2012**
Samsung electronics	1	4	2	1	3	2	4	6	5	5
Alphabet (Google)	2	1	1	6	6	11	13	26	46	44
Toyota Motor	3	3	3	2	1	1	2	8	6	4
Microsoft	4	5	8	7	11	13	16	22	25	28
PetroChina	5	2	4	3	5	3	1	1	1	2
Facebook	6	10	11	19	35	55	100	155	171	323
Volkswagen	7	7	7	5	4	6	8	10	9	11
China Mobile (2021)/China Telecom (2020-)	8	37	32	26	21	1533	27	41	57	53
Huawei investment & holding	9	13	15	16	20	42	64	93	103	104
Intel	10	9	9	8	10	14	12	15	12	12
Saudi Arabian Oil (Aramco)	11	6	5	-	-	-	-	-	-	-
General Motors	12	8	6	4	2	4	15	21	22	19
Apple	13	11	10	9	7	15	22	33	31	37
NTT	14	18	19	27	24	217	25	23	13	9
Taiwan Semiconductor	15	20	31	24	26	37	39	42	45	48
Kalera	16	-	-	-	-	-	-	-	-	-
China Petroleum & Chemical	17	16	24	40	36	18	14	12	10	8
Électricité de France	18	17	21	13	17	19	20	17	18	415
Exxon Mobil	19	12	14	15	14	7	5	7	3	2
Roche	20	23	23	21	22	24	35	39	37	31

Source: developed by the authors

Cells contain the ranking position of each company in the corresponding year. All companies in this list were founded before 2012. The newest one is Kalera, founded in 2010

Samples

Dataset A for the 2021 edition of the EC-IRI report includes 779 (31%) US companies, 597 (24%) Chinese companies, 401 (16%) EU companies, 293 (11%) Japanese companies, and 430 (17%) companies from the rest of the world. Dataset B for the 2021 edition of the NSF-HERD report includes 655 US universities. The numbers are nearly the same for all other editions of each report. Dataset A is formed by tuples with 19 items, including the company, the country, R&D, R&D 1-year-growth, and Capex (capital goods). Dataset B is formed by tuples with 13 items, including the university, all R&D expenditure, R&D expenditure on computer science, and R&D expenditure on information science. Both datasets were pre-processed (file name, column label, and data-type standardization) and analyzed for distribution, missing values, and outliers. Table 1 shows the top 20 R&D and Capex investors.

From dataset A, we chose R&D investments (EC_RD, henceforward) and Capex (EC_Capex, henceforward) as the items of interest. Missing values are usually handled by marginalization (discarding the missing values) or imputation (filling in missing values) (García-Laencina et al., 2010). The option for imputation implies choosing one of several strategies to obtain the value for each missing value. No missing values were found for EC_RD. Missing values found for EC_Capex (128 cases on average per year/report, or 5.1%) were filled with zero, considering that replacing it with, say, the mean or the median instead of zero would artificially amplify investments made in capital goods, with potential distortions on the clustering. Also, inference would not be suitable in this case because Capex does not depend on other factors in our dataset.

From dataset B, we chose all R&D expenditure (All_RD, henceforward) and R&D expenditure on computer and information sciences (CIS, henceforward) as the items of interest, considering that investments in R&D and CIS, especially made by tech companies, have been related to AI investments (Liu et al., 2021). No missing values were found for All_RD and CIS. Outliers were evaluated with the Mahalanobis distance (Riani et al., 2009) with a 0.001 significance level.

Dataset C includes 36,411 papers with data from AI conferences from 2013 to 2022. Dataset C is formed by tuples with four items: paper title, authors, affiliations, and funding. With an algorithm specifically developed for classification, each paper was classified according to four classes: (C1) authors' affiliation, with the following three categories: from universities, from companies, and from universities and companies; (C2) research funding, with the following seven categories: from universities, from companies, from government, from universities and companies, from universities and government, from companies and government, and from universities, government, and companies; (C3) a subclass of C1 when its content is "from university" in the form of a counter for the occurrence of each university according to Table 2; and (C4) a subclass of C1 when its content is "from company" in the form of a counter for the occurrence of each company according to Table 3. All algorithms developed for the classification of dataset C were implemented separately by two of the authors of this study, with a 98.9% inter-rater agreement. Table 4 shows the quantity of papers extracted for each conference between 2013 and 2022.

Results

Analysis of Datasets A and B and Test of Hypotheses H_1 and H_2

Figure 2 shows a leptokurtic shape for dataset A (which is also the case for dataset B). A leptokurtic shape "is a widely accepted" issue in financial data (Premaratne & Bera, 2005, p. 169). It includes Poisson, logistic, and student-*t* distributions, and it suggests the occurrence of outliers (Bossaerts, 2021). So, data distribution is abnormal and includes at least two data segments: data concentration and data dispersion (outliers). Further analysis on outliers shows that the current biggest R&D investors are companies like Alphabet, Huawei, Microsoft, Samsung, Apple, and Facebook.

Table 2 Dataset C (2013–2022)

Conference	**h-index**	**Number of papers**									
		2022	**2021**	**2020**	**2019**	**2018**	**2017**	**2016**	**2015**	**2014**	**2013**
CVPR: IEEE Conference on Computer Vision and Pattern Recognition	299	2074	1545	1466	1294	978	790	643	603	540	471
NeurIPS: Neural Information Processing Systems	198		2329	1905	1427	1009	679	569	403	411	360
ICCV: International Conference on Computer Vision	176		1582		1077		621		526		454
ACL: Annual Meeting of the Association for Computational Linguistics	135		572	794	753	533	246				
AAAI: Association for the Advancement of Artificial Intelligence	126	1269	1556	1558	1318						
EMNLP: Conference on Empirical Methods in Natural Language Processing	112		849	752	1302	1153					

Source: developed by the authors

Table 3 Clustering parameters and coefficients

Year	Dataset A			Dataset B		
	Clusters	SC	Eps	Clusters	SC	Eps
2012	9	.858	1.3	9	.826	1.1
2013	9	.881	1.68	9	.779	1.1
2014	9	.928	1.6	9	.784	.75
2015	9	.930	1.8	9	.775	.7
2016	9	.925	1.8	9	.779	1.2
2017	8	.932	1.7	8	.779	1.1
2018	9	.918	1.7	8	.762	1
2019	8	.936	2.5	9	.808	1.19
2020	9	.934	2.5	9	.792	1.19
2021	9	.936	2.5	8	.789	1.13

Source: developed by the authors

Figure 3 shows an example of cluster analysis applied to dataset A for the year 2021. The same analysis was done for each dataset and year. The scatter plot shows a region of high concentration at the bottom/left, and a region of dispersion at the top/right, with clusters in different colors.

The silhouette coefficient (SC) is a well-known measure of fitness for clustering (Pelka, 2018). SC assumes values from −1 to 1, and “[t]he higher value means the better assignment of objects into clusters” (Řezanková, 2018, p. 3). Table 5 shows the clustering parameters and coefficients for datasets A and B considering all years.

Table 4 Dataset A: clusters and companies (2021)

Cluster ID	Companies
#0	Alphabet
#1	Huawei Investments & Holding Apple
#2	Microsoft
#3	Samsung Electronics
#4	Facebook Volkswagen Intel
#5	Roche Johnson & Johnson Daimler + 4,485 other companies
#6	Toyota Motor
#7	PetroChina
#8	Saudi Arabian Oil China Mobile

Cluster ID is informed by the DBSCAN algorithm

Source: developed by the authors

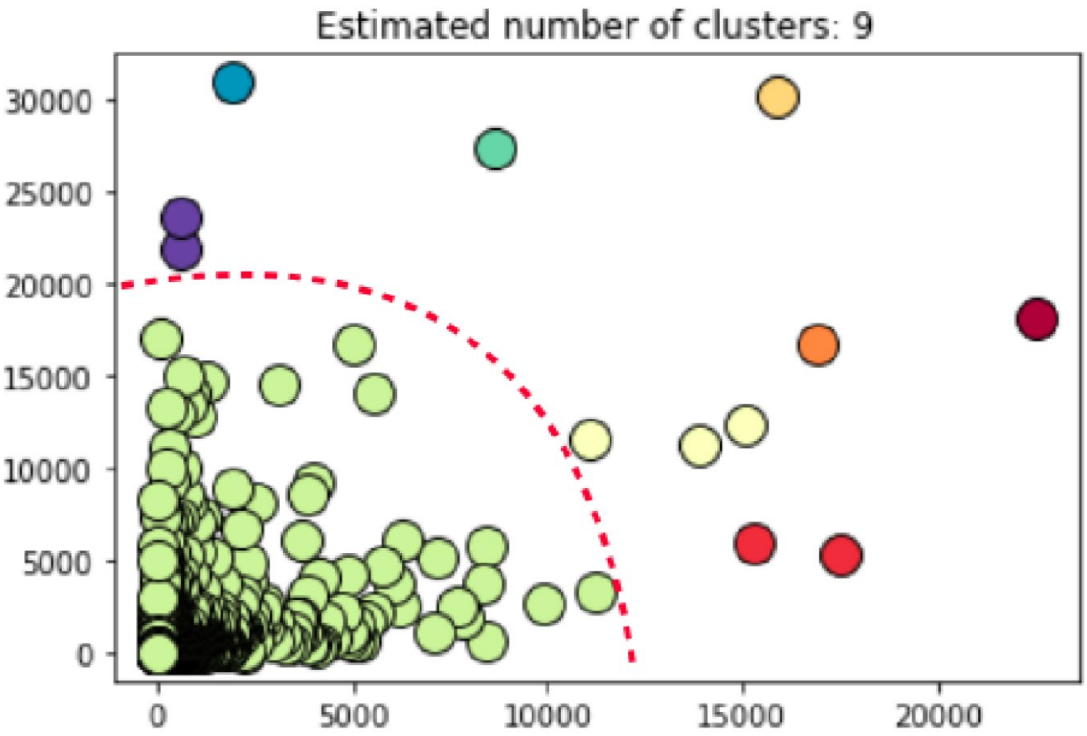


Fig. 3 Exemplary cluster analysis on dataset A (2021). Note: the dotted line suggests a threshold between ordinary investors and outliers Source: developed by the authors

Table 4 shows the clustering objects for dataset A, using data from 2021 to illustrate the clustering pattern we observed in all years. While 4488 companies form a single cluster (#5), 12 other companies form eight clusters.

Table 5 shows the clustering for dataset B, 2021. In this case, 646 universities form a single cluster, and nine other universities form eight additional clusters.

Table 5 Dataset B: clusters and universities (2021)

Cluster ID	Universities
#0	Johns Hopkins
#1	University of Michigan, Ann Arbor University of California, San Francisco University of Pennsylvania + 643 other universities
#2	University of Maryland, College Park
#3	Georgia Institute of Technology (GA Tech)
#4	Pennsylvania State University, University Park and Hershey Medical Center University of Illinois, Urbana-Champaign (UI)
#5	Massachusetts Institute of Technology (MIT) University of Southern California (USC)
#6	University of Texas, Austin (UT)
#7	Carnegie Mellon University (CMU)

Cluster ID is informed by the DBSCAN algorithm

Source: developed by the authors

Tables 6 and 7 respectively show the clustering from both datasets A and B in all years (2012–2021) obtained through the following steps, for dataset k, year y:

- First-order clustering using the algorithm DBSCAN has identified a number of c_{ky} clusters.
- Each cluster C is composed of j objects o, following the form $C_{k,y} = \{o_1, ..., o_\mathrm{j}\}$.
- Second-order clustering identified objects o_j often composing clusters c_{ky} (supposedly the ones investing heavily and permanently in R&D).

Table 6 shows the following:

1. There is a group of five companies that invest heavily and steadily in R&D and capital goods (Samsung Electronics, Intel, Toyota Motor, PetroChina, and Volkswagen). These are companies in the industries of “Software & Computer Services,” “Electronic & Electrical Equipment,” “Technology Hardware & Equipment,” and “Automobiles & Parts.”
2. There is a group of seven companies that have been gradually increasing investments in R&D and capital goods (Microsoft, Alphabet, Apple, Huawei, Aramco, Facebook, and China Mobile). These are companies in the “Software & Computer Services” and “Technology Hardware & Equipment” industries, with one company from the “Oil & Gas” industry.
3. There is a third group of 14 companies that have been gradually reducing investments in R&D and capital goods when compared to the other groups. This group consists mainly of companies in the “Oil & Gas,” “Automobiles & Parts,” and “Pharmaceuticals & Biotechnology” industries. The existence of these three second-order clusters suggests that there is an ongoing movement (in the larger R&D scenario) that can be explained by the occupation of empty spaces—companies occupying open investment spaces by companies from other industries (mostly “Oil & Gas”)—or an increased pressure from companies that “push” from other industries in the race for AI.

Table 7 shows the following:

1. There is a group of eight universities that expend heavily and steadily on R&D in general and on R&D related to CIS in particular (Johns Hopkins, Georgia Tech, Pennsylvania State University, MIT, USC, UT, UI, and CMU).
2. There is a second group of seven universities that have been gradually reducing expenditure on R&D and CIS when compared to the other groups. And unlike companies, there is no group of late-entrant universities (i.e., universities that have been gradually increasing expenditure on R&D).

To test H_1, we searched for a significant difference in investments in R&D and equipment (Capex) between groups of companies that invest more in R&D. To test H_2, we searched for a significant difference in expenditure in R&D between groups of universities that expend more in R&D. To assess both hypotheses, we analyzed each

Table 6 Dataset A: clusters and companies (2012–2021)

Company	Year									
	2021	2020	2019	2018	2017	2016	2015	2014	2013	2012
Samsung electronics	3	3	1	0	2	1	1	1	1	0
Intel	4	1	2	1	1	2	3	3	3	2
Toyota Motor	6	5	5	5	5	4	5	5	4	0
PetroChina	7	7	7	7	7	7	6	7	6	5
Volkswagen	4	1	2	1	0	0	0	0	0	2
Microsoft	2	1	2	1	1	2	2	2		
Alphabet (Google)	0	0	0	1	1	2	3			
Apple	1	1	2	1	4	5				
Huawei investment & holding	1	2	3	2						
Saudi Arabian Oil (Aramco)	8	8	7							
Facebook	4	1								
China Mobile	8									
General Motors		6	6	6	5	6				
Royal Dutch Shell				8	6	8	8	8	7	6
AT&T				8	6					
Daimler				4						
Exxon Mobil						8	8	8	7	7
Chevron						8	8	8	7	6
Gazprom						8		8	8	8
Total						8	7	8		
Petrobras							8	8		
Novartis								4		
Roche								4		
Nissan Motor									5	3
General Electric									5	3
NTT									5	4

Table 6 (continued)

Source: developed by the authors

Cells contain the ID of each cluster (ID informed by the DBSCAN algorithm). Colors represent "discretionary clusters" grouped by chronological assignment to DBSCAN clusters and quantity of assignments. "Year" is the year of data publication (fiscal year+1, where the fiscal year is a 1-year period that companies and governments use for reporting financial data)

year independently, thus forming 10 sub-hypotheses for each original hypothesis. Such a decision considers that objects o_j (companies and universities) often form distinct clusters c_{ky} with different R&D budget year after year. In order to form the groups for

Table 7 Dataset B: clusters and universities (2012–2021)

University	Year									
	2021	2020	2019	2018	2017	2016	2015	2014	2013	2012
Johns Hopkins	0	0	0	0	0	0	0	0	0	0
Georgia Institute of Technology (GA Tech)	3	4	4	4	4	6	6	7	6	5
Pennsylvania State University, University Park and Hershey Medical Center	4	5	3	3	3	5	5	5	5	4
Massachusetts Institute of Technology (MIT)	5	3	2	2	2	4	4	5	4	4
University of Southern California (USC)	5	6	5	4	5	7	6	7	6	6
University of Texas, Austin (UT)	6	5	6	5	5	5	7	7	6	5
University of Illinois, Urbana-Champaign (UI)	4	7	7	4	5	6	6	6	6	5
Carnegie Mellon University (CMU)	7	8	8	7	7	8	8	8	8	8
University of Maryland, College Park	2	2		6	6				7	7
University of California, San Francisco						2	3	4	3	3
University of California, San Diego						3		3	2	
University of Michigan, Ann Arbor							1	1		1
University of Washington, Seattle									2	
Ohio State University									5	
University of Chicago										7

Source: developed by the authors

Cells contain the ID of each cluster (ID informed by the DBSCAN algorithm). Colors represent "discretionary clusters" grouped by chronological assignment to DBSCAN clusters and quantity of assignments. "Year" is the year of data publication (fiscal year+1, where the fiscal year is a 1-year period that companies and governments use for reporting financial data)

Table 8 Datasets A and B, Mann-Whitney *U* tests for H_1 and H_2 (2012–2021)

Year	Dataset A			Dataset B		
	H_1	Statistic	*p*-value	H_2	Statistic	*p*-value
2012	$H_{1,a}$	29,279.0	< .001	$H_{2,a}$	10,552.0	< .001
2013	$H_{1,b}$	29,491.0	< .001	$H_{2,b}$	8131.0	< .001
2014	$H_{1,c}$	27,030.0	< .001	$H_{2,c}$	6815.0	< .001
2015	$H_{1,d}$	31,978.0	< .001	$H_{2,d}$	6053.0	< .001
2016	$H_{1,e}$	27,074.0	< .001	$H_{2,e}$	6107.0	< .001
2017	$H_{1,f}$	33,833.0	< .001	$H_{2,f}$	5451.0	< .001
2018	$H_{1,g}$	29,150.0	< .001	$H_{2,g}$	5486.0	< .001
2019	$H_{1,h}$	33,868.0	< .001	$H_{2,h}$	4909.0	< .001
2020	$H_{1,i}$	23,092.0	< .001	$H_{2,i}$	5536.0	< .001
2021	$H_{1,j}$	17,146.5	< .001	$H_{2,j}$	5610.0	< .001

Source: developed by the authors

comparison (i.e., to assess the difference between two samples), groups were formed by objects that are above and below the threshold between ordinary investors and outliers, as illustrated in Fig. 3. The procedure was repeated for each dataset and each year. Table 8 shows that hypotheses H_1 and H_2 found statistical support.

Analysis of Dataset C and Test of Hypotheses H_3 and H_4

We have hypothesized that the role of universities in the production and exploitation of firm-specific knowledge is peripheral from the perspective of the final product design, production, and marketing, and it occurs because (1) companies seek profit and competitive advantage, whereas universities focus on research rather than attempting to market products and services; and (2) even when universities focus on entrepreneurial activities mainly through the creation of spin-off firms, such spin-offs are often bought by companies. So, to test hypotheses H_3 and H_4, we analyzed the academic output (from dataset C) of LC (clustered from dataset A) and LU (clustered from dataset B). Table 9 shows the quantity of papers published (the academic output) by authors affiliated to LC from 2013 to 2022.

From the 36,411 papers in dataset C (Table 9), 5742 (15.7%) have authors affiliated with LC. Of those, 2107 (36.7%) include a statement on funding, with 33.5% of funding coming from government and universities, which suggests that LC research funding is mostly internal (66.5%). Table 10 additionally shows that 711 papers were co-authored with LU. This seems to be not significant, but when we consider the collaboration with any university, we find 4161 papers. That is, 72.4% (about three-fourths) of papers published by the 12 LC are co-authored with universities.

To test H_3, we searched for a significant difference between the sum of papers authored by LC (last column of Table 9) and the quantity of co-authorships between LC and LU (penultimate column of Table 10). To test H_4, we searched for a significant difference between the sum of papers authored by LC (last column of Table 9) and the quantity of co-authorships between LC and any

Table 9 Database C: quantity of papers authored by LC (2013–2022)

Company	**Number of papers**										
	2022	**2021**	**2020**	**2019**	**2018**	**2017**	**2016**	**2015**	**2014**	**2013**	**Total**
Alphabet/Google	109	463	410	384	242	125	67	63	33	27	1,923
Microsoft	107	360	311	307	159	116	56	94	55	80	1,645
Apple	10	15	15	22	12	2	0	1	0	0	77
Huawei	108	217	115	84	23	2	0	2	8	5	564
Samsung	35	62	55	61	16	3	2	3	0	5	242
Facebook	33	257	174	229	88	42	12	18	4	3	860
Intel	18	44	33	64	33	28	12	8	3	7	250
Volkswagen	0	3	0	3	4	0	0	0	0	1	11
Toyota	15	22	20	38	18	10	11	5	4	13	156
Petro China	0	1	0	0	1	0	0	0	0	0	2
China Mobile	0	3	6	0	0	0	0	0	0	0	9
Saudi Arabian Oil/Aramco	0	2	0	1	0	0	0	0	0	0	3
Total	435	1449	1139	1193	596	328	160	194	107	141	5742

Source: developed by the authors

university (last column of Table 10). Firm-specific AI knowledge was measured as the quantity of papers published by authors affiliated to LC. To assess H_3, the statistical test considered the null hypothesis as the absence of any difference between the volume of academic production from the LC and the academic output from LC in collaboration with LU. Similarly, for H_4, the statistical test posited the null hypothesis as the absence of any difference between the volume of academic output from the LC and the academic output from LC in collaboration with universities in general. At a 95% confidence level, the application of the test indicated that we cannot reject the null hypotheses (respectively $p = 0.052$ and $p = 0.582$) in either scenario. Since the null hypotheses were not rejected in both instances, we have evidence to support the theoretical hypotheses H_3 and H_4, suggesting an intense partnership between LC and universities in general in producing firm-specific AI knowledge. Moreover, such an intense partnership between LC and universities in general is corroborated by a 72.4% rate of papers they have co-authored.

Table 10 also shows that companies more closely related to IT publish much more than the auto and oil ones at AI conferences. That is, R&D investments of the IT companies are positively correlated with the quantity of papers published in the main AI conferences. This is evidence in favor of using R&D investments as a proxy for AI investments, in the case of IT companies.

Table 10 Co-authorships between LC, LU, and any university (2013–2022)

Company	**Number of papers co-authored between LC and LU**										**Number of papers co-authored between LC and any university**
	JH	**UM**	**GIT**	**PSU**	**UI**	**MIT**	**USC**	**UT**	**CM**	**Total**	
Alphabet/Google	24	24	19	4	15	28	14	19	78	225	1204
Microsoft	14	27	12	4	38	8	4	40	71	218	1309
Apple	1	1	0	1	0	1	0	5	5	14	43
Huawei	11	3	0	4	1	1	1	3	3	27	509
Samsung	1	0	5	0	3	0	0	4	0	13	161
Facebook	21	13	35	1	14	4	8	25	53	174	656
Intel	0	4	3	0	3	1	0	4	6	21	173
Volkswagen	0	0	0	0	0	0	0	0	0	0	8
Toyota	0	0	0	0	4	1	1	2	10	18	87
Petro China	0	0	0	0	0	0	0	0	0	0	1
China Mobile	0	0	0	0	0	0	0	0	0	0	8
Saudi Arabian Oil/Aramco	0	0	0	0	0	0	0	1	0	1	2
Sum	72	72	74	14	78	44	28	103	226	711	4161

Source: developed by the authors

JH Johns Hopkins, *UM* University of Maryland, *GIT* Georgia Institute of Technology, *PSU* Pennsylvania State University, *UI* University of Illinois, *MIT* Massachusetts Institute of Technology, *USC* University of Southern California, *UT* University of Texas, *CM* Carnegie Mellon

Discussion

We proposed a rationale for the over-concentration of AI knowledge (OCAIK) as a consequence of the investment power of a few companies. In the proposal, the relationship between OCAIK and investments in R&D and capital goods is supported by the partnership with universities. The proposal foresees social, technical, ethical, and political impacts from such relationships. We consider that the model applies to contexts of competitiveness, in which companies that have slack resources seek to gain competitive advantages through the development of firm-specific knowledge.

Based on related works (Benetta et al., 2021; Yiu et al., 2020), we assume that investments by companies in R&D and capital goods (Capex) are suitable proxies for investments in AI. This seems to be a novel and useful approach. Even the number of patents or registration of new products is not a better variable to measure AI investments. For example, the AI Watch technical report from the European Commission Joint Research Centre uses OCDE data "to monitor the development, uptake and impact of AI for Europe" (Benetta et al., 2021, p. 1), but the results are an estimate as there are difficulties in computing patents and AI-related training (adopted as AI investment proxies). Our study, in its turn, found that companies more closely related to IT are much more involved in AI conference papers than companies from other industries. This is evidence in favor of using R&D as a proxy for AI in the case of IT companies. This study also differs from others in that we use institution-level data, whereas related work uses data aggregated by country, region, or industry.

We found a significant difference between clusters of companies that we now call *ordinary investors* and *outlier investors* (H_1). In 2021, for example, we found 4488 companies forming a single cluster (#5), while 12 other companies formed eight different clusters. This is the default pattern found in other years as well in dataset A. This finding supports the proposed rationale, especially in the relationship between OCAIK, R&D, and capital goods investments. This is also in line with Marimon and Quadrini (2011) in that the dominance over an industry favors the over-concentration of knowledge, as well as in line with Arnosti and Weinberg (2022) in that the ownership of the most efficient resources interposes barriers to competition.

We also found support for H_2, i.e., there is a significant difference in expenditure on R&D between groups of universities that expend more in R&D. In 2021, for example, we found 646 universities comprising a single cluster, while nine other universities form eight different clusters. This is the default pattern for other years as well in dataset B. This finding supports the proposed rationale, especially about the role of universities in the relationship between OCAIK, R&D, and capital goods investments, which is also in line with Choi and Contractor (2019) who state that due to complexity and research costs, companies turn to partnerships in R&D, and also in line with Mahdi et al. (2019), who posit that partnerships occur preferably with universities.

Regarding the role of the partnership of universities on the relationship between R&D and capital goods investment, and LC firm-specific AI knowledge, we did not find a significant difference between the academic output of LC integrated with LU and the academic output of LC integrated with universities in general (including LU), from which we conclude that there is an intense partnership between LC and universities in the production of firm-specific knowledge, which supports H_3 and H_4. In fact, about three-fourths (72.4%) of papers published by 12 LC are co-authored with universities.

Such findings suggest the existence of OCAIK represented by a few companies with a high capacity for investment in association with universities. OCAIK can be illustrated with a singularity metaphor, like in the case of a black hole, where density and gravity are so high that even light cannot escape the force of attraction at the point of no return. Our metaphor resembles a differentiation singularity: an insurmountable distance—at least during a certain period of time—between those who compete and those who are beyond the limits of competition. In this scenario, a group of players is considered singular when the sum of firm-specific knowledge already available by a firm added to its capacity to invest in new firm-specific knowledge development surpasses the sum of publicly available knowledge on that subject added to the capacity of other players to invest in firm-specific knowledge. Business competition may thus reach a point of no return if only incremental innovation is devised instead of radical innovation (Lindbloom, 1959). The expression below synthesizes the idea:

(e1) S_{it} if $(\mathrm{SKA}_{\mathrm{it}} + \mathrm{NSK}_{\mathrm{it}}) > (\sum \mathrm{PK} + \sum_1^n \mathrm{NSK})$
S, differentiation singularity
i, firm index
t, specific period of time
SKA, firm-specific knowledge already available
NSK, capacity for new firm-specific knowledge development
PK, publicly available knowledge on that subject
n, number of other players

Theoretical Implications

This study makes a number of contributions to the RBV and KBV literature. First, it points out an apparent limitation of both theories: how much advantage is needed to keep a sustainable advantage, while still keeping it competitive? To the best of our knowledge, both theories have no answer to that question. Second, the study proposes a metaphor for OCAIK, that of a differentiation singularity, to be considered in RBV and KBV. Metaphors and other cognitive resources are useful to illustrate situations and help identify key issues to be studied, designed, and managed (Jácome de Moura & Porto-Bellini, 2019). From the idea of a differentiation singularity, we may conceive the existence of a business oligopoly able to dictate the rules of the AI industry without being threatened by any real competitor. A third contribution of this study is to avoid investigating the performance of over-concentrated knowledge sectors in the light of RBV, KBV, or any other theory of the firm that assumes the existence of competition. That is, this study proposes a limit for theoretical explanations and distinguishes itself from recent systematic literature reviews on RBV and KBV that point out future trends in the development of theories (e.g., Pereira & Bamel, 2021) while not setting theoretical limits. This study additionally contributes to the literature on AI governance and regulation as it offers empirical evidence for the existence of OCAIK, the role of universities, and the potential consequences of the dominance of only a few industry actors over the creation and the evolution of AI knowledge.

Managerial Implications

Our findings show that companies that master the state-of-the-art in AI have distanced themselves from others beyond possible competition, since they are likely to have better resources, investment capacity, or motivation to work closely with scholarly research. Policy makers must be aware of this fact when planning ways to stimulate both the development and the regulation of AI, since the already established business capacity, organizational networks, and market presence may be far more difficult to manage than one can devise. Similarly, other companies should also be aware of that if they plan to compete in the AI arena through mere incremental innovation.

Moreover, our findings show that about three-fourths of papers published by 12 LC are co-authored with universities. This fact has two implications. First, managers of companies interested in AI-related R&D can benefit from an already open channel of interaction with universities. And second, policymakers and university managers must be cognizant of the resources being allocated to fund scholarly research, with such funding possibly interfering in business competition.

Limitations and Future Studies

The first limitation of this study refers to the databases included in the analyses. Even if we took measures to include all the relevant sources, there is always the possibility that important sources were ignored. Another limitation is that the coding procedure was highly dependent on the authors' discretion, even if including independent coding and formal assessments of agreement between two of the authors. Another possible limitation involves the use of R&D as a proxy for AI. This limitation is also present in the literature, since even the number of patents or registration of new products are no better variables to measure AI investments (Benetta et al., 2021). And another limitation is that we have restricted our analysis to data showing the presence of companies and universities in AI scholarly conferences. While we have employed quality criteria to select the sources of data, it is inevitable to think of data selection bias, once papers in conferences represent but a fraction of what is under development in the partnership of businesses and the academic institutions, not to say in the full AI landscape.

As for future studies, besides addressing some of the limitations above, we suggest that researchers could measure the dominance of the global AI industry through companies' investments specifically focused on AI development. Such a metric should be obtained through specific regulation that requires the disclosure of data on investments in AI, as AI regulation is a trendy topic in the global agenda.

Conclusion

This study discussed OCAIK as a manifestation of the dominance over the global AI industry by a small number of companies that are able to articulate with resourceful universities the development and the use of cutting edge, inimitable AI knowledge. Scholarly research is thus done in the periphery of business and, in planned or fortuitous manners, feeds business with research findings that grant those companies with increasing competitive advantages over new entrants. We verify the occurrence of OCAIK by processing thousands of papers presented in AI academic conferences in the last decade. Besides, we offer a novel insight about the RBV and KBV theoretical traditions, in that business competition may reach a point of no return if only incremental innovation is devised instead of radical innovation to break the chains of knowledge accumulation and technological implementation by a limited number of agents. By presenting evidence of a business oligopoly able to dictate the rules of the AI industry without being threatened by any real competitor, this study thus conveys political, economic, and social implications, since such evidence—i.e., obtained through close monitoring of companies' behavior—is the first step to counteract oligopolistic tendencies in the AI industry. We suggest monitoring the AI industry through a permanent panel/observatory, which could be championed by research institutions or organizations such as the OECD or WTO, for example.

Appendix. Synthesis of the Procedures for Mining, Collecting, and Processing the Three Datasets

Dataset	Link	Procedure
A	https://iri.jrc.ec.europa.eu/scoreboard	Step 1: access the main link Step 2: access specific link for each year Step 3: download the data (in CSV format) Step 4: remove empty columns in each file, as well as descriptive headers and totaling lines Step 5: standardize column/variable names Step 6: data processed through IDE Spider Python v3.9.7 using scipy and sklearn libraries
B	https://www.nsf.gov/statistics/srvyherd/#tabs-2	Step 1: access the main link Step 2: access specific link for each year Step 3: download the data (in CSV format) Step 4: remove empty columns in each file, as well as descriptive headers and totaling lines Step 5: standardize column/variable names Step 6: data processed through IDE Spider Python v3.9.7 using scipy and sklearn libraries

Dataset	Link	Procedure
C	Link provided according to each conference. Examples: https://openaccess.thecvf.com/CVPR2021 https://proceedings.neurips.cc/paper/2021 https://openaccess.thecvf.com/ICCV2021 https://aclanthology.org/volumes/2021.acl-long/ https://aaai.org/Library/AAAI/aaai21-issue02.php#4 https://aclanthology.org/volumes/2021.emnlp-main/	Step 1: web scraping to access each PDF file published by each conference Step 2: extract the paper titles, authors, affiliations, and funding information from each PDF file, composing a new CSV file Step 3: classify each record (in the new CSV file) according to: 3.1: go through CSV, line by line 3.2: check if affiliations have any company 3.3: check if CSV affiliations have any universities 3.4: check if funding has any company 3.5: check if funding has any university 3.6: check if funding has any government agency 3.7: if the paper has affiliations from universities, then C1 = 1 3.8: if the paper has affiliations from companies, then C1 = 2 3.9: if the paper has affiliations from universities and companies, then C1 = 3 3.10: if the paper has funding from universities, then C2 = 1 3.11: if the paper has funding from companies, then C2 = 2 3.12: if the paper has funding from the government, then C2 = 3 3.13: if the paper has funding from universities and companies, then C2 = 4 3.14: if the paper has funding from universities and government, then C2 = 5 3.15: if the paper has funding from companies and government, then C2 = 6 3.16: if the paper has funding from universities, government, and companies, then C2 = 7 Step 4: data processed through IDE Spider Python v3.9.7 using scipy library Step 5: 98.9% inter-rater agreement between two independent developers

Data Availability Data will be made available upon request. No artificial intelligence tool was used in this study.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Arnosti, N., & Weinberg, S. M. (2022). Bitcoin: A natural oligopoly. *Management Science, 68*(7), 4755–4771. https://doi.org/10.1287/mnsc.2021.4095

Assael, Y., Sommerschield, T., Shillingford, B., et al. (2022). Restoring and attributing ancient texts using deep neural networks. *Nature, 603*(7900), 280–283. https://doi.org/10.1038/s41586-022-04448-z

Bahoo, S., Cucculelli, M., & Qamar, D. (2023). Artificial intelligence and corporate innovation: A review and research agenda. *Technological Forecasting & Social Change, 188*, 122264. https://doi.org/10.1016/j.techfore.2022.122264

Bai, X., & Li, J. (2020). The best configuration of collaborative knowledge innovation management from the perspective of artificial intelligence. *Knowledge Management Research & Practice*, 1–13. https://doi.org/10.1080/14778238.2020.1834886

Barney, J. (1991). Firm resources and sustained competitive advantage. *Journal of Management, 17*(1), 99–120. https://doi.org/10.1177/014920639101700108

Benetta, A. D., Sobolewski, M., & Nepelski, D. (2021). *AI Watch: 2020 EU AI investments* (No. JRC126477). Joint Research Centre. https://doi.org/10.2760/017514

Bharadwaj, A. S. (2000). A resource-based perspective on information technology capability and firm performance: An empirical investigation. *MIS Quarterly, 24*(1), 169–196. https://doi.org/10.2307/3250983

Boiko, K. (2021). R&D activity and firm performance: Mapping the field. *Management Review Quarterly, 71*(1), 1–37. https://doi.org/10.1007/s11301-021-00220-1

Bossaerts, P. (2021). How neurobiology elucidates the role of emotions in financial decision-making. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2021.697375

Braune, C., Besecke, S., & Kruse, R. (2015). Density based clustering: Alternatives to DBSCAN. In *Partitional Clustering Algorithms* (pp. 193–213). Springer, Cham. https://doi.org/10.1007/978-3-319-09259-1_6

Bruderer, H. (2016). The birth of artificial intelligence: First conference on artificial intelligence in paris in 1951? In: *IFIP International Conference on the History of Computing*, 181–185. https://doi.org/10.1007/978-3-319-49463-0_12

Bui, X. N., Nguyen, H., Choi, Y., Nguyen-Thoi, T., Zhou, J., & Dou, J. (2020). Prediction of slope failure in open-pit mines using a novel hybrid artificial intelligence model based on decision tree and evolution algorithm. *Scientific Reports, 10*(1), 1–17. https://doi.org/10.1038/s41598-020-66904-y

Cai, Y., Ramis Ferrer, B., & Luis Martinez Lastra, J. (2019). Building university-industry co-innovation networks in transnational innovation ecosystems: Towards a transdisciplinary approach of integrating social sciences and artificial intelligence. *Sustainability, 11*(17), 4633. https://doi.org/10.3390/su11174633

Castelvecchi, D. (2015). Artificial intelligence called in to tackle LHC data deluge. *Nature, 528*(7580), 18–19. https://doi.org/10.1038/528018a

Challa, H., Niu, N., & Johnson, R. (2020). Faulty requirements made valuable: On the role of data quality in deep learning. In *2020 IEEE Seventh International Workshop on Artificial Intelligence for Requirements Engineering (AIRE),* 61–69. https://doi.org/10.1109/AIRE51212.2020.00016

Choi, J., & Contractor, F. J. (2019). Improving the progress of research & development (R&D) projects by selecting an optimal alliance structure and partner type. *British Journal of Management, 30*(4), 791–809. https://doi.org/10.1111/1467-8551.12267

Cohen, J. F., & Olsen, K. (2013). The impacts of complementary information technology resources on the service-profit chain and competitive performance of South African hospitality firms. *International Journal of Hospitality Management, 34*, 245–254. https://doi.org/10.1016/j.ijhm.2013.04.005

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*. https://doi.org/10.1016/j.ijinfomgt.2021.102383

D’Este, P., & Perkmann, M. (2010). Why do academics engage with industry? The entrepreneurial university and individual motivations. *The Journal of Technology Transfer, 36*(3), 316–339. https://doi.org/10.1007/s10961-010-9153-z

Dafoe, A. (2018). *AI governance: A research agenda*. Oxford, UK, Future of Humanity Institute, University of Oxford.

Dahiya, R., Le, S., Ring, J. K., & Watson, K. (2021). Big data analytics and competitive advantage: The strategic role of firm-specific knowledge. *Journal of Strategy & Management*. https://doi.org/10.1108/jsma-08-2020-0203

Degravel, D. (2011). Managing organizational capabilities: The keystone step. *Journal of Strategy and Management, 4*(3), 251–274. https://doi.org/10.1108/17554251111152270

Ding, J., & Dafoe, A. (2021). The logic of strategic assets: From oil to AI. *Security Studies, 30*(2), 182–212. https://doi.org/10.1080/09636412.2021.1915583

Edwards, C. (2021). The best of NLP. *Communications of the ACM, 64*(4), 9–11. https://doi.org/10.1145/3449049

Eisenhardt, K. M., & Martin, J. A. (2000). Dynamic capabilities: What are they? *Strategic Management Journal, 21*(10–11), 1105–1121. https://doi.org/10.1002/1097-0266(200010/11)21:10/11%3c1105::AID-SMJ133%3e3.0.CO;2-E

Etzkowitz, H., & Leydesdorff, L. (1995). The triple helix—university-industry-government relations: A laboratory for knowledge-based economic development. *EASST Review*, *14*(1), 14–19. Available at SSRN: https://ssrn.com/abstract=2480085

Ferreira, J. J., Fayolle, A., Ratten, V., & Raposo, M. (2018). *Introduction: The role of entrepreneurial universities in society*. Edward Elgar Publishing. https://doi.org/10.4337/9781786432469.00005

Fink, L. (2011). How do IT capabilities create strategic value? Toward greater integration of insights from reductionistic and holistic approaches. *European Journal of Information Systems, 20*(1), 16–33. https://doi.org/10.1057/ejis.2010.53

Formica, P. (2002). Entrepreneurial universities: The value of education in encouraging entrepreneurship. *Industry & Higher Education, 16*(3), 167–175. https://doi.org/10.5367/000000002101296261

Fredström, A., Wincent, J., Sjödin, D., Oghazi, P., & Parida, V. (2021). Tracking innovation diffusion: AI analysis of large-scale patent data towards an agenda for further research. *Technological Forecasting & Social Change*. https://doi.org/10.1016/j.techfore.2020.120524

García-Laencina, P. J., Sancho-Gómez, J. L., & Figueiras-Vidal, A. R. (2010). Pattern classification with missing data: A review. *Neural Computing & Applications, 19*(2), 263–282. https://doi.org/10.1007/s00521-009-0295-6

Ge, S., & Liu, X. (2022). The role of knowledge creation, absorption and acquisition in determining national competitive advantage. *Technovation, 112*, 102396. https://doi.org/10.1016/j.technovation.2021.102396

Grant, R. M. (1996). Toward a knowledge-based theory of the firm. *Strategic Management Journal, 17*(S2), 109–122. https://doi.org/10.1002/smj.4250171110

Grassano, N., Hernandez, H., Fako, P., Tuebke, A., Amoroso, S., Georgakaki, A., ... & Panzica, R. (2022). *The 2021 EU Industrial R&D Investment Scoreboard*, No. JRC127360. Joint Research Centre. https://doi.org/10.2760/559391

Hamet, P., & Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism, 69*(Supplement), S36–S40. https://doi.org/10.1016/j.metabol.2017.01.011

Hassanzadeh, P., Atyabi, F., & Dinarvand, R. (2019). The significance of artificial intelligence in drug delivery system design. *Advanced Drug Delivery Reviews, 151*, 169–190. https://doi.org/10.1016/j.addr.2019.05.001

Henard, D. H., & McFadyen, M. A. (2006). *R&D knowledge is power. Research-Technology Management, 49*(3), 41–47. https://doi.org/10.1080/08956308.2006.11657377

Hou, J., Gao, H., & Li, X. (2016). Dsets-DBSCAN: A parameter-free clustering algorithm. *IEEE Transactions on Image Processing, 25*(7), 3182–3193. https://doi.org/10.1109/TIP.2016.2559803

Hu, Y., Kuang, W., Qin, Z., Li, K., Zhang, J., Gao, Y., & Li, K. (2021). Artificial intelligence security: Threats and countermeasures. *ACM Computing Surveys (CSUR), 55*(1), 1–36. https://doi.org/10.1145/3487890

Huggins, R., Johnston, A., & Stride, C. (2012). Knowledge networks and universities: Locational and organisational aspects of knowledge transfer interactions. *Entrepreneurship & Regional Development, 24*(7–8), 475–502. https://doi.org/10.1080/08985626.2011.618192

Jácome de Moura, P., Jr., & Porto-Bellini, C. G. (2019). Shared flow in teams: Team vibration as emergent property, metaphor, and surrogate measure. *Team Performance Management, 25*(7/8), 440–456. https://doi.org/10.1108/TPM-12-2018-0072

Jeon, S., Chang, Y. S., & Jo, S. J. (2024). Speed of catch-up and convergence of the artificial intelligence divide: AI investment, robotic, start-ups, and patents. *Journal of Global Information Technology Management*, *27*(1). https://doi.org/10.1080/1097198X.2023.2297636

Johnston, A. (2019). The roles of universities in knowledge-based urban development: A critical review. *International Journal of Knowledge-Based Development, 10*(3), 213–231. https://doi.org/10.1504/IJKBD.2019.103205

Karami, A., & Johansson, R. (2014). Choosing DBSCAN parameters automatically using differentiation evolution. *International Journal of Computer Applications, 91*(7), 1–11.

Kaur, D., Uslu, S., & Durresi, A. (2020). Requirements for trustworthy artificial intelligence—a review. In *International Conference on Network-Based Information Systems* (pp. 105–115). https://doi.org/10.1007/978-3-030-57811-4_11

Kerr, A., Barry, M., & Kelleher, J. D. (2020). Expectations of artificial intelligence and the performativity of ethics: Implications for communication governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951720915

Kovacevich, A. (2022). *Take it from a software engineer: Big tech's monopoly is stifling innovation*. Accessed June 28, 2022. https://www.newsweek.com/take-it-software-engineer-big-techs-monopoly-stifling-innovation-opinion-1718646

Lai, W., Zhou, M., Hu, F., Bian, K., & Song, Q. (2019). A new DBSCAN parameters determination method based on improved MVO. *IEEE Access, 7*, 104085–104095. https://doi.org/10.1109/ACCESS.2019.2931334

Li, K., Zhang, J., & Li, D. (2021). Research status and hotspot analysis of literature metrology in artificial intelligence field. *Journal of Physics: Conference Series, 2024*(1), 012055. https://doi.org/10.1088/1742-6596/2024/1/012055

Li, Z., Li, Y., Lu, W., & Huang, J. (2020). Crowdsourcing logistics pricing optimization model based on DBSCAN clustering algorithm. *IEEE Access, 8*, 92615–92626. https://doi.org/10.1109/ACCESS.2020.2995063

Lindbloom, C. E. (1959). The science of "muddling through." *Public Administration Review, 19*(2), 79–88.

Liu, N., Shapira, P., & Yue, X. (2021). Tracking developments in artificial intelligence research: Constructing and applying a new search strategy. *Scientometrics, 126*(4), 3153–3192. https://doi.org/10.1007/s11192-021-03868-4

Mahdi, O. R., Nassar, I. A., & Almsafir, M. K. (2019). Knowledge management processes and sustainable competitive advantage: An empirical examination in private universities. *Journal of Business Research, 94*, 320–334. https://doi.org/10.1016/j.jbusres.2018.02.013

Marakova, V., Wolak-Tuzimek, A., & Tuckova, Z. (2021). Corporate social responsibility as a source of competitive advantage in large enterprises. *Journal of Competitiveness, 13*(1), 113–128. https://doi.org/10.7441/joc.2021.01.07

Marimon, R., & Quadrini, V. (2011). Competition, human capital and income inequality with limited commitment. *Journal of Economic Theory, 146*(3), 976–1008. https://doi.org/10.1016/j.jet.2011.01.001

Martin, D., Fowlkes, C., Tal, D., & Malik, J. (2001). A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: *Proceedings 8th IEEE International Conference on Computer Vision. ICCV 2001*, 416–423. https://doi.org/10.1109/ICCV.2001.937655

Mayer, K. J., Somaya, D., & Williamson, I. O. (2012). Firm-specific, industry-specific, and occupational human capital and the sourcing of knowledge work. *Organization Science, 23*(5), 1311–1329. https://doi.org/10.1287/orsc.1110.0722

Menke, M. M. (1997). Managing R&D for competitive advantage. *Research-Technology Management, 40*(6), 40–42. https://doi.org/10.1080/08956308.1997.11671169

Mikhaylov, S. J., Esteve, M., & Campion, A. (2018). Artificial intelligence for the public sector: Opportunities and challenges of cross-sector collaboration. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2128). https://doi.org/10.1098/rsta.2017.0357

Miotti, L., & Sachwald, F. (2003). Co-operative R&D: Why and with whom? An integrated framework of analysis. *Research Policy, 32*(8), 1481–1499. https://doi.org/10.1016/S0048-7333(02)00159-2

Miotto, G., Del-Castillo-Feito, C., & Blanco-González, A. (2020). Reputation and legitimacy: Key factors for higher education institutions' sustained competitive advantage. *Journal of Business Research, 112*, 342–353. https://doi.org/10.1016/j.jbusres.2019.11.076

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill Science/Engineering.

Mongeau, S., Hajdasinski, A. (2021). Managerial recommendations. In: *Cybersecurity Data Science*. Springer, Cham. https://doi.org/10.1007/978-3-030-74896-8_6

Nayak, B., Bhattacharyya, S. S., & Krishnamoorthy, B. (2022). Exploring the black box of competitive advantage—an integrated bibliometric and chronological literature review approach. *Journal of Business Research, 139*, 964–982. https://doi.org/10.1016/j.jbusres.2021.10.047

Newbert, S. L. (2008). Value, rareness, competitive advantage, and performance: A conceptual-level empirical investigation of the resource-based view of the firm. *Strategic Management Journal, 29*(7), 745–768. https://doi.org/10.1002/smj.686

Newman, D. (2017). *Inside look: The world's largest tech companies are making massive AI investments*. Accessed May 4, 2022. https://www.forbes.com/sites/danielnewman/2017/01/17/inside-look-the-worlds-largest-tech-companies-are-making-massive-ai-investments/?sh=4d7f17cc4af2

OECD. (2020). *A first look at the OECD's framework for the classification of AI systems, designed to give policymakers clarity*. Accessed May 2, 2022. https://oecd.ai/en/wonk/a-first-look-at-the-oecds-framework-for-the-classification-of-ai-systems-for-policymakers

Orlando, B., Ballestra, L. V., Magni, D., & Ciampi, F. (2021). Open innovation and patenting activity in health care. *Journal of Intellectual Capital, 22*(2), 384–402. https://doi.org/10.1108/JIC-03-2020-0076

Pang, B., Nijkamp, E., & Wu, Y. N. (2020). Deep learning with TensorFlow: A review. *Journal of Educational & Behavioral Statistics, 45*(2), 227–248. https://doi.org/10.3102/1076998619872761

Parmigiani, F., Haller, I., Gkantsidis, C., & Ballani, H. (2021, May). Optics for the cloud: Challenges and opportunities. In: *CLEO: Science and Innovations* (pp. STu1J-2). Optical Society of America.

Pełka, M. (2018). Analysis of innovations in the European Union via ensemble symbolic density clustering. *Ekonometria*, *22*(3). https://doi.org/10.15611/eada.2018.3.06

Pereira, V., & Bamel, U. (2021). Extending the resource and knowledge-based view: A critical analysis into its theoretical evolution and future research directions. *Journal of Business Research, 132*, 557–570. https://doi.org/10.1016/j.jbusres.2021.04.021

Premaratne, G., & Bera, A. (2005). A test for symmetry with leptokurtic financial data. *Journal of Financial Econometrics, 3*(2), 169–187. https://doi.org/10.1093/jjfinec/nbi009

Ramanathan, V., Wang, R., & Mahajan, D. (2021). Predet: Large-scale weakly supervised pre-training for detection. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2865–2875. https://doi.org/10.1109/ICCV48922.2021.00286

Řezanková, H. A. N. A. (2018). Different approaches to the silhouette coefficient calculation in cluster evaluation. In: *21st International Scientific Conference AMSE Applications of Mathematics and Statistics in Economics*, 1–10.

Riani, M., Atkinson, A. C., & Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society: Series B, 71*(2), 447–466. https://doi.org/10.1111/j.1467-9868.2008.00692.x

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & society, 36*(1), 59–77. https://doi.org/10.1007/s00146-020-00992-2

Rosenblatt, F. (1960). Perceptron simulation experiments. *Proceedings of the IRE, 48*(3), 301–309. https://doi.org/10.1109/JRPROC.1960.287598

Samuel, A. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research & Development., 44*(1), 206–226. https://doi.org/10.1147/rd.33.0210

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS), 42*(3), 1–21. https://doi.org/10.1145/3068335

Shannon, C. E. (1950). Programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 41*(314), 256–275. https://doi.org/10.1080/14786445008521796

Shao, Z., Yuan, S., & Wang, Y. (2020). Institutional collaboration and competition in artificial intelligence. *IEEE Access, 8*, 69734–69741. https://doi.org/10.1109/ACCESS.2020.2986383

Smuha, N. A. (2021). From a 'race to AI'to a 'race to AI regulation': Regulatory competition for artificial intelligence. *Law, Innovation & Technology, 13*(1), 57–84. https://doi.org/10.1080/17579961.2021.1898300

Soni, N., & Ganatra, A. (2016). Aged (automatic generation of eps for DBSCAN). *International Journal of Computer Science & Information Security, 14*(5), 536.

Statista. (2022). Estimated number of companies worldwide from 2000 to 2020. *Statista*. Accessed March 28, 2022. https://www.statista.com/statistics/1260686/global-companies/#statisticContainer

Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. *Strategic Management Journal, 18*(7), 509–533. https://doi.org/10.1002/(SICI)1097-0266(199708)18:7%3c509::AID-SMJ882%3e3.0.CO;2-Z

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Uhr, L., & Vossler, C. (1961). A pattern recognition program that generates, evaluates, and adjusts its own operators. In: *Papers presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference*, 555–569. https://doi.org/10.1145/1460690.1460751

Ullah, Z., & Arslan, A. (2022). R&D contribution to sustainable product attributes development: The complementarity of human capital. *Sustainable Development, 30*(5), 902–915. https://doi.org/10.1002/sd.2289

Wang, C., Chen, M. N., & Chang, C. H. (2020). The double-edged effect of knowledge search on innovation generations. *European Journal of Innovation Management, 23*(1), 156–176. https://doi.org/10.1108/EJIM-04-2018-0072

Wang, H., Choi, J., Wan, G., & Dong, J. Q. (2016). Slack resources and the rent-generating potential of firm-specific knowledge. *Journal of Management, 42*(2), 500–523. https://doi.org/10.1177/0149206313484519

Webb, A. (2019). *The big nine: How the tech titans and their thinking machines could warp humanity*. NY, PublicAffairs.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45. https://doi.org/10.1145/365153.365168

Yiu, L. D., Yeung, A. C., & Jong, A. P. (2020). Business intelligence systems and operational capability: An empirical analysis of high-tech sectors. *Industrial Management & Data Systems, 120*(6), 1195–1215. https://doi.org/10.1108/IMDS-12-2019-0659

Zahra, S. A., Neubaum, D. O., & Hayton, J. (2020). What do we know about knowledge integration: Fusing micro-and macro-organizational perspectives. *Academy of Management Annals, 14*(1), 160–194. https://doi.org/10.5465/annals.2017.0093

Zhao, B. (2017). Web scraping. *Encyclopedia of big data*, 1–3.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Pedro Jácome de Moura Jr.[1] · **Carlos Denner dos Santos Junior**[2] · **Carlo Gabriel Porto-Bellini**[1] · **José Jorge Lima Dias Junior**[1]

✉ Pedro Jácome de Moura Jr.
pjacome@sti.ufpb.br

Carlos Denner dos Santos Junior
carlosdenner@unb.br

Carlo Gabriel Porto-Bellini
cgpbellini@ccsa.ufpb.br

José Jorge Lima Dias Junior
jorge.dias@academico.ufpb.br

[1] Department of Management, UFPB, João Pessoa, Brazil

[2] Department of Management, UnB, Brasilia, Brazil