NOVA IMS Information Management School

MEGI

Master Degree Program in

Statistics and Information Management

AI Ethics Guidelines

Is regulation a value of trust?

Ana Patrícia Poças Pires Bulha Almeida

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

AI Ethics Guidelines
Is Regulation a value of trust?

By

Ana Patricia Poças Pires Bulha Almeida

Master Thesis presented as partial requirement for obtaining the master's degree in Statistics and Information Management, with a specialization in Risk Analysis and Management

Supervised by

Bruno Miguel Pinto Damásio, PhD, NOVA Information Management School
and Sandro Miguel Ferreira Mendonça, PhD, ISCTE-IUL

July, 2025

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism, any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisboa, July 2025]

Ana Patrícia Poças Pires Bulha Almeida

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to all those who supported me throughout the journey of my thesis.

First, I am sincerely thankful to my parents and Luis for their unwavering support, encouragement, and love. To my five siblings, Catarina, Georgette, João, Juliana, and Valter, thank you for always being there, in your own unique ways, cheering me on.

To my grandmother, whose quiet strength and affection have been a source of comfort, and to all my friends, who stood by me through both challenges and achievements, your presence has meant more than words can express.

To Lulu and Benji, my feline companions, whose silent presence, warmth, and gentle affection were a constant source of comfort, even without words.

I am also grateful to my director, Dra. Margarida, and to my entire work team for their continuous motivation and collaboration, which contributed greatly to my academic and personal development during this process.

Finally, I extend my heartfelt thanks to my supervisors, Professor Bruno and Professor Sandro, for their guidance, insightful feedback, and generosity with their time and knowledge. Their support has been invaluable to the completion of this work.

ABSTRACT

This research investigates the ethical challenges associated with Artificial Intelligence (AI), focusing on the role of regulation as a potential enabler of trust. It addresses the central research question: Can regulation be considered a value of trust in AI? Based on a comparative analysis of existing AI ethical guidelines and empirical findings from experts, this study investigated which legal frameworks support ethical principles guidelines. The research adopts a mixed-method approach, combining bibliometric analysis, literature review, and normative evaluation. Findings suggest that regulation not only reinforces ethical compliance but also serves as a stabilizing mechanism that fosters legitimacy and confidence in AI systems. This study clarifies the connection between ethics, regulation, and trust, on AI systems, by identifying key ethical values most aligned with current regulatory mechanisms.

KEYWORDS

Artificial Intelligence; Ethic; Regulation; Transparency; Accountability; Trust; Guidelines; Governance; Algorithm; Responsible.

Sustainable Development Goals (SDG):

TABLE OF CONTENTS

1.Introduction **1**

2. Literarure Review **3**

2.1 AI Ethics 3

2.2 AI Ethics Guidelines 4

2.2.1 Transparency 6

2.2.1.1 The “Black Box” Problem 7

2.2.2 Explicability 8

2.2.3 Justice, Fairness, and Equity 9

2.2.3.1 Social impact 10

2.2.4 Non-Maleficence 11

2.2.5 Responsibility And Accountability 12

2.2.6 Privacy 13

2.3 Relevance of AI Ethic Guidelines 14

2.3.1 Perception and Relevance of AI Ethics Principles: A Comparative Analysis 14

2.3.2 Ethic AI Guidelines as a Governance Tool? 15

2.3.3 From Principles To Procedure 16

2.4 Trustworthy AI 17

2.4.1 How to Implement Trustworthy AI 18

2.4.2 The Legalization of Ethic Guidelines 20

3. Methodology **22**

3.1 Structure of Methodological Approach 22

3.2 Comparative Analysis of Ethical Guidelines 25

4. Results **31**

4.1 Data Presentation 31

4.1.1 Frequency in the Literature (Jobin et Al., 2019) 31

4.1.2 Expert Evaluation (Rothenberg & Fabian, 2019) 32

4.1.3 Combined Score (Literature + Experts) 32

4.1.4 Final Score Including Regulatory Alignment 33

4.2 Discussion on the Results 33

5. Conclusions and Future Research **35**

6. Bibliographical References 37

LIST OF FIGURES

Figure 1 - From Principles to Practice Flow 17
Figure 2 - Research Framework 22
Figure 3 - Keywords extracted from VOS Viewer 25

LIST OF TABLES

Table 1 - Ethical principles identified in existing AI guidelines 5
Table 2 - Results of the filtering process 23
Table 3 - Selected Ethical Guidelines 26
Table 4 - Ranking of Ethical Guidelines by the Experts 27
Table 5 - Rescaled Scores and Ranking of Ethical Guidelines Based on Expert Evaluation 28
Table 6 - Combined Scores and Ranking of Ethical Guidelines (Jobin and Expert Evaluation) 28
Table 7 - Regulatory score 29
Table 8 - Final Ranking of Ethical Guidelines 30
Table 9 – Frequency in institucional documents (Jobin et. al., 2019) 32
Table 10 – Expert evaluation (Rothenberg & Fabian, 2019) 32
Table 11 – Combined score (Literature + Experts) 32
Table 12 – Final score including regulatory alignment 33

LIST OF ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
AI HLEG	High Level Expert Group on Artificial Intelligence
EC	European Commission
EU	European Union
GDPR	General Data Protection Regulation
GTM	Grounded Theory Methodology
OECD	Organization for Economic Co-operation and Development
RAI	Responsible AI
WOS	Web of Science

1. INTRODUCTION

The increasing integration of Artificial Intelligence (AI) into critical sectors such as healthcare, justice, finance, public administration, and security has raised some ethical concerns, not only limited to the technical aspects of how AI systems make decisions, but to their implications for human rights, social justice, cultural values, and institutional accountability.

In response to these concerns, a range of initiatives, both public and private, have sought to develop ethical guidelines to guide the responsible design, deployment, and governance of AI technologies. Notably, the study by Jobin et al. (2019) provides a comprehensive review of 84 AI ethics frameworks, identifying five core principles consistently emphasized across contexts: transparency, responsibility, non-maleficence, justice and fairness, and privacy. Despite this convergence at the level of principles, their practical implementation remains fragmented and is often not accompanied by binding regulatory mechanisms.

Complementing this normative perspective, Rothenberger and Fabian (2019) adopt an empirical approach, combining expert interviews with public opinion surveys to evaluate which ethical principles are most valued by diverse stakeholders. Their findings offer a valuable bridge between the academic discourse on AI ethics and the expectations held by society.

This study adopts a mixed-method research that includes bibliometric filtering, comparative policy analysis, and normative evaluation. It seeks to identify the ethical principles most aligned with existing regulatory frameworks and to evaluate how effectively regulation can serve as a stabilizing and trustworthy force within AI governance.

The present dissertation explore the role of regulation as a value, specifically, whether it can be conceptualized as a foundational component in building and maintaining trust within the ethical landscape of AI.

The structure of this dissertation is as follows. Chapter 2 presents a comprehensive literature review on AI ethics, with a focus on the development of ethical guidelines, their relevance, and their potential for governance mechanisms. It explores core principles such as transparency, justice, non-maleficence, responsibility, and privacy, as well as the challenges of achieving trustworthy AI.

Chapter 3 details the methodology adopted for empirical analysis, including a bibliometric review and a comparative evaluation of ethical frameworks.

Chapter 4 outlines and discusses the findings, combining data from institutional documents and expert assessments to assess how ethical principles align with regulatory instruments.

Finally, Chapter 5 concludes the study by summarizing key insights, acknowledging its limitations, and outlining directions for future research.

2. LITERATURE REVIEW

The growing digitalization of artificial intelligence (AI) has profoundly transformed our society. In this context, rapid technological progress and its integration into our daily lives have raised some ethical concerns about the social impact of these new technologies.

In recent years, IA has had a tremendous impact on every field and has brought huge economic and social benefits (Lu & Xu, 2019). But recently some ethical challenges have been reported during the development of intelligent systems (Saghiri, Vahidipour, Jabbarpour, Sookhak & Forestiero, 2022).

This literature review will analyze the evolution of various approaches and perspectives, from different countries, on the ethical and social issues associated with the subject under analysis, as well as the role of regulation as a promoter of trust and transparency in the progressive technological development. In this sense, to provide a more in-depth understanding of the topic, some case studies will be presented, which reflect the challenges associated with incorporating the ethical metrics under analysis, and some guidelines that may lead to future research in the area.

2.1 AI ETHICS

The study of AI ethics involves a systematic exploration of moral reasoning – namely, what is considered right or wrong, just or unjust – in the context of AI development. Several scholars have highlighted the increasing importance of this field, particularly considering the need to align AI technologies with fundamental ethical principles and societal values (Floridi et al., 2018; Jobin, Lenca & Vayena, 2019). These studies emphasize the necessity for comprehensive ethical analysis, not only to address conceptual ambiguities but also to confront practical challenges in the implementation of AI RAI (Responsible AI) systems.

AI ethics is a field that has emerged as a response to the increasing concerns about the impact of artificial intelligence. Kazim and Koshiyama (2021) highlight that the development of this field reflects the need to address the ethical challenges associated with the expanding role of AI in society. Their overview underscores how AI ethics has evolved to critically assess these impacts, facilitating the establishment of frameworks that guide its responsible integration into various domains (Kazim & Koshiyama, 2021).

2.2 AI ETHICS GUIDELINES

The current exponential growth of artificial intelligence (AI) has been accompanied by an increasing demand for the application of ethics, aiming to harness the disruptive potential of emerging technologies while minimizing associated risks. In this context, the establishment and prioritization of ethical guidelines have become essential—not only to provide safeguards, but also to foster public trust and acceptance of AI systems. These guidelines typically include normative principles and practical recommendations intended to guide the responsible development and use of AI RAI (Hagendorff, 2020).

As a result, a considerable number of ethical frameworks have been developed in recent years (Hagendorff, 2020). In his work, Hagendorff (2020) critically examines the consistency and limitations of many of these documents, while Larsson (2021) discusses their role as governance tools, particularly within the context of the European Union. According to a widely cited study by Jobin, Lenca, and Vayena (2019; Table 1), at least 84 public and private initiatives had published ethical AI guidelines since the mid-2010s (Larsson, 2021). Notably, 88% of these documents were released after 2016, reflecting an intensification of ethical concerns during a key period in AI's development. Transparency was identified as a key principle in 73 of the 84 sources analyzed.

The study indicates that 88% of these guidelines were published post 2016, with transparency highlighted as a critical principle in 84 sources (Larsson, 2021). Larsson (2021) also highlights that these initiatives span a wide range of actors, from major technology companies such as Google and Telia, to research centers like the AI Now Institute, as well as governmental and intergovernmental organizations, including the AI HLEG (High Level Expert Group on Artificial Intelligence**)** of the European Commission, the OECD's (Organization for Economic Co-operation and Development) AI in Society group, Singapore's Advisory Council on the Ethical Use of AI and Data, and the UK House of Lords Select Committee on Artificial Intelligence.

Jobin et al. (2019) concluded that there is a global consensus around at least five core ethical principles, including transparency, justice and fairness, non-maleficence, responsibility, and privacy. These principles are referenced in over half of the review sources. However, further thematic analysis reveals significant and conceptual variation in how the eleven ethical principles are interpreted and in the specific recommendations or areas of concern that emerge from each other.

Table 1 - Ethical principles identified in existing AI guidelines

Ethical Principle	Number of documents	Included codes
Transparency	73/84	Transparency, explain ability, explicability, understandability, interpretability, communicative, disclosure and showing
Justice & Fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non)bias, (non)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental) and non-subversion
Responsibility	60/84	Responsibility, accountability, liability and acting with integrity
Solidarity	6/84	Solidarity, social security and cohesion
Privacy	47/84	Privacy and personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good and common good
Freedom & Autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty and empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy and resources (energy)
Dignity	13/84	Dignity

Note. Adapted from Jobin, Lenca & Vayena, 2019.

A comprehensive evaluation of these core principles, their conceptual foundations, practical implications, and ongoing challenges, will be presented in the following sections, beginning with the principle of transparency.

2.2.1 TRANSPARENCY

Transparency is the most frequently cited principle in the present literature, referenced in 73 out of 84 references (Jobin et al., 2019). Despite this apparent consensus, there is considerable variation in this concept´s meanings, rationales, application areas, and methods of execution. According to the author, references regarding transparency generally include initiatives to improve explicability, interpretability, and other disclosure and communication channels.

The principle of transparency is particularly relevant in the context of AI regulation. It plays a central role in enabling effective governance, serving as a precondition for accountability. (Abrassart, et. al, 2018). As Abrassart et.al emphasize, AI systems should be designed and implemented in a way that allows for continuous supervision and scrutiny. This includes making all phases of the system's lifecycle—design, development, and deployment—accessible and understandable for relevant stakeholders (Fjeld et al., 2020). Without sufficient transparency, regulators and affected individuals alike face significant challenges in identifying risks, assigning responsibility, or seeking redress.

The most relevant application domains include data use, human-AI interaction, automated decisions and the motivation behind data use or application of AI systems (Jobin et al., 2019). But also, some sources related transparency to democratic values, debate and involvement, legal reasons, and its role in building trust (Jobin et al., 2019).

This principle states that AI systems should be implemented in a way that permits their oversight during operations (Abrassart, et. al, 2018). Throughout the AI system lifecycle, the stages of design, development, and implementation should be accessible for examination (Fjeld et al., 2020).

The Toronto Declaration states that adequate transparency cannot be guaranteed, particularly in high-risk contexts. This guidance reinforces the notion that transparency is not merely a technical feature, but a legal and ethical obligation in certain regulatory environments (Amnesty International, 2018). Jobin et al. (2019) note that many ethical frameworks associate transparency not only with system performance and accountability, but also with fostering public trust and legal compliance. In this sense, transparency functions as a foundational condition for the responsible and legitimate governance of AI technologies.

2.2.1.1 THE “BLACK BOX” PROBLEM

Pasquale (2015) states that algorithms are considered "black boxes," non-transparent software tools. These are used in the development of AI software by specialists who may lack a comprehensive understanding of their functioning and the sensitivity to recognize the magnitude of their impact.

As required in the GDPR (General Data Protection Regulation) (2016), users have the right to access relevant information regarding automated decisions, often referred to as the "right to explanation." This right includes knowledge about system functionality, specific decisions, and their rationale, whether ex-ante or ex-post (Edwards & Veale, 2017; Selbst & Powles, 2018; Wachter, Mittelstadt & Floridi, 2016).

However, according to Domingos (2015) and Olhede & Wolfe (2018), some authors argue that providing detailed explanations about the functioning of an algorithm or AI system, or about the rationale behind an automated decision, may be undesirable. From this perspective, such systems can advance various fields more efficiently than humans, thus making explanations less of a priority.

The issue of accountability in the use of algorithms has increasingly concerned politicians and policymakers. One example is the initiative of former German Chancellor Angela Merkel, who called for the development of platforms that inform users about the utility and purpose of their data (Council of Europe, 2017).

Although it is not feasible to ensure complete transparency of the code used due to intellectual property rights, the Council of Europe (2017) advocates for the disclosure of other information, such as the variables employed, training data, values and deviations, as well as the quantity and type of data processed. This approach aims to achieve what is referred to as “effective transparency.” Moreover, it is important to emphasize that the data feeding automated systems are considered more relevant than the algorithms themselves (Council of Europe, 2017).

The European Commission (2018) further emphasizes that auditing AI systems is essential, as their evaluation by internal and external auditors, along with the availability of the resulting reports, promotes trust in such technology.

O'Neil (2016) also highlights the necessity of auditing algorithms. According to Kroll et al. (2017), auditing algorithms involve verifying whether they comply with specific requirements. Such audits should be conducted through collaboration between governments

and civil society organizations. However, Villani (2018) emphasizes that intellectual property rights represent a significant obstacle to public auditing algorithms.

Regarding transparency in AI governance, Almeida (2021) argues that ensuring transparency does not necessarily imply that citizens need to understand the code or its functionality. Nevertheless, according to Weng (2021), the public shows interest in engaging in discussions on this subject, aiming to share their concerns, particularly regarding the safe use of government services (e.g., e-governance).

For Almeida et al. (2021), the concept of transparency is associated with the decision-making process, particularly the norms guiding the use of technology, the origin of the data used, and the risks associated with AI systems.

In this context, citizens prefer that the public sector provide clear information regarding the impact of AI on daily life, explaining how this technology can improve public services and whether its implementation aligns with ethical values (Wirtz et al., 2022). Weng et al. (2021) suggest that public organizations adopt modern communication platforms, such as social networks (Twitter, Facebook, among others), to disseminate information more effectively and facilitate interaction between organizations and citizens. According to the authors, this approach could bridge the communication gap between governments and citizens, thereby promoting greater transparency. According to Morley et al. (2020), the application of algorithms in such contexts carries the risk of compromising public trust.

2.2.2 EXPLICABILITY

According to Fjeld and colleagues (2020), the principles of transparency and explicability are the most frequently cited principles in review documents, appearing in approximately three-quarters of the analyzed documents.

As stated by Kazim and Koshiyama (2021), transparency is associated with the principle of openness, which is fundamental to the ethics of AI systems, as it establishes trust and accountability in their use. In this context, transparency includes two main aspects: the decisions regarding how AI systems are utilized and the way these systems reach their conclusions. The first aspect is directly related to governance, while the second pertains to explicability, that is, the capacity to make the processes leading to automated decisions comprehensible (Abrassart et al., 2018; Fjeld et al., 2020).

Explicability is particularly relevant for addressing the “black box” problem, enabling AI systems to be clearly evaluated by different stakeholders. This evaluation varies depending

on the technical expertise of the individuals involved, their roles in the system's lifecycle, and their interaction with the system. To ensure explicability, technical requirements and tools are necessary to trace and monitor decisions, providing either global explanations (related to the model) or specific ones (focused on individual data points) (Kazim & Koshiyama, 2021). Therefore, translating technical concepts and decision outcomes into comprehensible and verifiable formats becomes a priority (Fjeld et al., 2020).

The importance of explicability is emphasized in various contexts, particularly in systems that significantly impact individuals' lives or reputations or affect their quality of life (Abrassart et al., 2018). An AI system that has a substantial impact on an individual's life should not be implemented if it cannot provide a complete and satisfactory explanation of its decisions (UK House of Lords, 2018). The Toronto Declaration reinforces that explicability is a prerequisite for impact assessments, establishing accountability and holding involved parties responsible (Amnesty International, 2018). Similarly, the European Commission associates explicability with the principle of non-discrimination, emphasizing that the development of comprehensible AI is essential to minimize the risks of bias and errors (European Commission, 2018).

In addition to explicability, clear communication about the capabilities and purposes of AI systems is indispensable. For instance, in systems that imitate human subjectivity (e.g., chatbots), it is crucial to inform users that they are interacting with an AI system (Kazim & Koshiyama, 2021).

2.2.3 JUSTICE, FAIRNESS, AND EQUITY

Ethics in AI faced significant challenges, with one of the main issues being the presence of racial (Chouldechova, 2017) and gender biases (Caliskan et al., 2017). These biases are evident in AI systems, which perpetuate social inequalities and diminish public trust in these technologies (Zuiderveen Borgesius, 2020).

According to Jobin and colleagues (2019), justice is often understood as a matter of equity, prevention, and mitigation of biases and discrimination. However, some sources also associate the concept with respect for diversity, inclusion, and equality, while other approaches emphasize the need to establish mechanisms that allow individuals to contest automated decisions and access processes for redress and remedy, ensuring the possibility of correcting unfair or harmful decisions. Furthermore, this study highlights that justice in AI is also related

to equitable access to data and the benefits provided by these technologies (Jobin et al., 2019). To preserve justice in AI, Jobin and colleagues (2019) suggest the following strategies:

1. **Technical solutions:** adoption of standards or explicit normative coding of principles.
2. **Transparency:** provision of information and raising public awareness of existing rights and regulations.
3. **Testing, monitoring, and auditing:** solutions promoted by data protection authorities.
4. **Strengthening the rule of law:** ensuring the right to appeal, recourse, and redress.
5. **Systemic and procedural changes:** governmental actions, inclusion of diverse disciplinary of demographic perspectives in teams, and active involvement of civil society and other relevant stakeholders in the design and oversight of AI systems.

As previously mentioned, algorithms can act as enablers of discrimination. Therefore, it is essential to establish robust governance and regulatory processes (Eyert et al., 2022; Yoo & Lai, 2020 to protect individuals from systemic discrimination caused by algorithms (Ulnicane et al., 2021). Ulmicane et al. (2021) argue that a governance framework for AI must incorporate rules that prevent these algorithm-driven discriminations.

Several studies provide concrete examples of discrimination associated with the use of algorithms. These include Sweeney's (2013) investigation, which identified discrimination in online advertisements; Buolamwini and Gebru's (2018) study, which revealed biases in facial recognition systems favoring lighter skin tones; Veale and Binns' (2017) analysis, which uncovered racial and social biases in the use of residential location to infer ethnicity or socioeconomic status; and Pandey and Caliskan's (2021) investigation, which identified unequal impacts in pricing algorithms used by transportation applications, resulting in discriminatory practices.

2.2.3.1 SOCIAL IMPACT

In this section, a case study is presented to exemplify the ethical and social implications of using AI in recruitment processes. The case under analysis is based on the work of Gupta and Mishra (2022), which provides a comprehensive overview of the challenges faced by organizations when implementing AI-driven recruitment tools.

To ensure ethical and responsible use, AI technologies must be developed with the integration of fundamental moral values and ethical principles. The rapid global expansion of

AI has given rise to an increasing number of ethical concerns that require critical examination and appropriate regulatory responses. (Grupta & Mishra, 2022)

In recent years, organizations have adopted a range of AI-based tools, including chatbots, facial recognition software, and automated screening algorithms, to support their recruitment processes. (Gupta & Mishra, 2022).

The authors emphasize the importance of ensuring that AI recruitment systems must function equitably, prevent discriminatory patterns, and protect candidates' privacy and data.

According to the study, the most prominent ethical issues in this domain are data privacy and unconscious bias. Gupta and Mishra (2022) explain that such bias results from the fact that AI algorithms are built using data provided by humans, who may reflect the unconscious attitudes and assumptions of those who created them.

Consequently, AI systems may unintentionally reproduce social inequalities, raising questions about accountability and transparency in automated decision-making systems.

Moreover, gender disparities in the AI sector contribute to stereotypes within AI systems. According to a UNESCO report cited by Gupta and Mishra (2022), only 22% of AI professionals worldwide are women. As example is the default feminization of virtual assistants such as Siri or Alexa, which are frequently programmed with female voices and submissive characteristic. This study indicates that AI has the potential to reinforce traditional gender roles and contribute to inequalities in society.

2.2.4 NON-MALEFICENCE

According to the analysis by Jobin and colleagues (2019), references to the concept of 'non-maliciousness' cover appeals to safety and security, as well as claims that AI should not cause foreseeable or unintended damage. According to sources detailing this analysis, to avoid possible risks or damage, such as cyber-attacks or malicious hacking, a robust risk management strategy must be adopted.

Associated with this theme, the concept of 'harm' is interpreted as an effect of discrimination, violation of privacy or physical damage. The author also points out that, in the literature analyzed, this term is less often associated with a loss of trust in AI systems, 'radical individualism', the risk associated with technological progress being faster than the ability to develop and implement adequate regulatory measures, negative impacts on social well-being in

the long term, and psychological, emotional and economic aspects. The author suggests the following damage prevention guidelines:

a) **Technical solutions:** assessments of data quality or security and privacy by design.
b) **Governance strategies:** proposals that include active co-operation, stakeholders, compliance with existing or new legislation, and the need to adopt supervisory practices (testing, monitoring, audits and evaluations).

The study by Jobin et al (2019) also reveals that most of the sources analyzed mention the concept of 'dual use', i.e., the possibility of technology being used both for beneficial purposes and to cause harm, which can be avoided if there has been a risk assessment, mitigation measures and a clear attribution of responsibility.

2.2.5 Responsibility And Accountability

The study by Jobin et al. (2019) reveals that the concept of responsibility and accountability is rarely defined in literature, despite frequent approaches to RAI. The actors identified as responsible for the actions or decisions of AI systems are creators, designers, institutions or industry (Jobin et al., 2019)

Some authors argue that responsibility should be aligned with rights, human values (Council of Europe, 2017; European Group of Ethics, 2018) and integrity (Jobin et al. 2019). Responsibility is linked to accountability and is considered a necessary condition for the social acceptability of AI (Council of Europe, 2017; European Commission, 2018; Rathenau Institute, 2017; Villani, 2018). But one of the most pressing questions in this area is whether AI should be held responsible for its actions in the same analogue way as humans, or whether humans should be solely responsible, since they are the ones who develop, implement and use it (Jobin et al., 2019). Or even whether this responsibility should be shared (Coeckelbergh, 2016). The Council of Europe (2017) states that it is difficult to assign legal or political responsibility to any of these parties since their level of control is unclear.

For the other hand, Chiao (2019) defends that holding AI responsible for the algorithm's actions or decisions should not be equivalent to holding human beings responsible. According to his study, the 'traditional' legal system, an area run by judges and lawyers, does not have professionals with the technical skills to assess whether the algorithm's behavior is unethical or not. The person responsible for a situation of damage is therefore the human being (software programmers or public administrators (Chiao, 2019).

According to the European Commission (2018), mechanisms must be created to guarantee responsibility and accountability for AI systems and their results, prior to their implementation on the market. Almeida et al. (2021) state that the EU, unlike the US, has a governance model, including policy guidelines, which hold programmers and public organizations accountable for the unethical use of AI.

Organizations and programmers who develop and implement AI systems must comply with specific procedures, such as drawing up documentation and adopting safety and security measures (Almeida et al., 2021; Matus & Veale, 2022). The existence of more autonomous systems and the automation of decisions makes it difficult to assign responsibilities and accountability (Rathenau Institute, 2017). According to this author, some documents (e.g. Villani, 2018; European Group on Ethics, 2018) state that researchers, developers, programmers and organizations must act responsibly and legally (Villani, 2018). The 'right to explanation', mentioned earlier in this analysis, could help to make people more responsible and accountable (Rathenau Institute, 2017). Singh, Cobbe, & Norval (2018) state that better approaches to this issue are being developed.

2.2.6 PRIVACY

In the context of AI ethics, privacy is considered a fundamental right that must be safeguarded, as it is associated with data protection, security, as well as freedom and trust in AI systems (Jobin et al., 2019). This theme represents a significant concern for most individuals, who pursue to preserve their individuality in an increasingly complex digital environment (Du & Xie, 2021; Kololakis, 2017). However, a paradox arises between the desire for personal privacy and the vast amounts of data shared online, often without full awareness of the associated risks (Kokolakis, 2017).

This paradox is particularly evident in the context of AI governance, where privacy is challenged by continuous access to digital services. To mitigate these risks, regulatory approaches have been proposed, including legal compliance certification models and, the adaptation of laws to the specificities of AI, aiming to promote trust and security in AI systems (Jobin et al., 2019). However, technological advancements that have simplified access to services such as e-government have also brought consequences, including a lack of privacy and the absence of robust guarantees for data protection (Wirtz et al., 2020; Willems et al., 2022).

Experts in the field have raised concerns about security gaps between personal devices, AI capabilities, and the processing of personal information. These risks are particularly pronounced in contexts such as the use of tablets, smartphones, or public devices to access governmental services, exposing personal data to potential threats (Willems et al., 2022; Du & Xie, 2021).

In a hyper connected world, "smart" devices, toys, refrigerators, or mobile phones are equipped with AI systems that continuously store and transmit data to manufacturers, amplifying potential threats. Pauwels (2020) further suggests that professionals responsible for the development, implementation, and maintenance of such AI systems should receive specialized training to strengthen data protection and security.

Proposals to address these challenges are typically categorized into three main approaches: technical solutions, including differential privacy, privacy by design, data minimization, and access controls; increased investment in research and public awareness; and the implementation of regulatory measures (Jobin et al., 2019). Nevertheless, governmental agencies continue to face challenges in ensuring the security of shared personal information (Wirtz et al., 2020). Kazim and Koshiyama (2021) also propose that informed consent is essential to ensure individuals are aware of how their data is used and stored.

The GDPR, which has been applied since 2018 in the European Union, establishes strict standards for the protection of personal data processed by algorithms. Although it's full effectiveness has yet to be observed, GDPR has already influenced legislation outside Europe, such as the California Consumer Privacy Act, from 2018. In this context, most of the documents that address this matter emphasize the value of developing common algorithms (Abrassart et al., 2018), open research and collaboration to support the advancement of technology. These regulations reflect the growing need for robust legal frameworks to protect individual privacy rights, as noted by Garcia et al. (2022).

2.3 Relevance of AI Ethic Guidelines

2.3.1 Perception and Relevance of AI Ethics Principles: A Comparative Analysis

Following the analysis previously presented of the study by Jobin et al. (2019) which identified thematic convergence around five core ethical principles, it becomes pertinent to compare these findings with other works that adopt complementary empirical approaches.

A notable example is the study by Rothenberger and Fabian (2019), which employed a mixed-methods methodology to explore the relevance attributed to ethical guidelines by both experts and citizens. Drawing on sources from academic, industrial, governmental, and associative contexts, the authors applied the Grounded Theory Methodology (GTM) during the initial qualitative phase, which involved seven interviews with experts (both internal and external to Information and Communication Technology organizations). In the second phase, an online survey was conducted with 51 citizens, aimed at validating and quantifying the perceptions gathered. Based on these two data sets, a weighted arithmetic mean was calculated, allowing the ethical principles to be ranked according to their perceived importance.

The articulation between the study by Jobin et al. (2019) and that of Rothenberger and Fabian (2019), highlights point of convergence and complementarity between documentary analysis and the social and professional perception of ethical guidelines. This approach enhances the understanding of the practical relevance of the identified principles, going beyond their mere presence in institutional documents.

Given the framework of this dissertation, which focuses on the relationship between ethics and regulation, it is proposed to introduce a third dimension of analysis: the degree of alignment between the ethical principles and the main legal and regulatory instruments currently in force or under discussion, namely the GDPR (2016), the White Paper on Artificial Intelligence: A European Approach to Excellence and Trust (European Commission, 2020) and the Toronto Declaration (Amnesty Internacional & Access Now, 2018).

2.3.2 ETHIC AI GUIDELINES AS A GOVERNANCE TOOL?

Despite the growing adoption of ethical guidelines in the field of artificial intelligence, their effectiveness as tools of governance remains a matter of debate**.** Several authors question whether these guidelines truly the capacity to guide organizational behavior or whether they function merely as symbolic instruments.

Hagendorff (2020) points out that one of the problems with the ethical guidelines developed in this area is that it lacks mechanisms to reinforce its own normative claims. The author criticizes the strategic use of ethical guidelines by AI companies and research institutes, which incorporate their ethical considerations and ‘self-commitments’ into the formulation of ethical guidelines on the use of AI. It also emphasizes that this method is an ‘evasion of regulation’ and that many companies ‘self-regulate’ to avoid stricter external regulation (Larsson, 2020).

Mark Coechelbergh, a member of the European Commission AI HLEG, outlines his concerns about ethical issues and the regulatory challenges of AI. He warns that there is a risk of ethics being used in a superficial way, like a 'fig leaf', i.e. that it can help ensure the acceptability of the technology and economic gain but then has no consequences for those who misuse these technologies (Larsson, 2020).

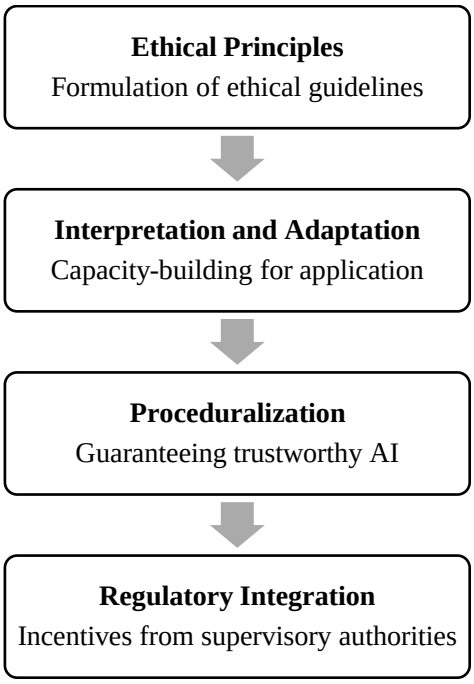
On the other hand, Larsson (2020) suggest that the limitations of ethical guidelines may not only stem from corporate strategy, but also from the complexity and speed of technological development**,** which challenge the timely creation of effective legislation. In this view, ethics may serve as an intermediate framework, providing a reference point while waiting for more critical research to inform the creation of solid legal foundations. Nevertheless, the question remains as to which elements of AI ethics are suitable for legal codification**,** with areas such as human oversight identified as particularly relevant (Koulu, 2020).

2.3.3 From Principles To Procedure

The set of ethical guidelines for the development and implementation of AI is based on principles but lacks procedural provisions for its implementation. So, the formulation of principles is the first, necessary phase. This is followed by the procedural phase, in which both the capacity to apply these principles needs to be strengthened and reliable AI needs to be guaranteed for all citizens, public authorities, consumers and for companies looking for incentives to invest in this type of system (Larsson, 2021).

This process is ilusrrated in Figure 1, which outline the flow from ethical principle formulation to regulatory integration.

Figure 1 - From Principles to Practice Flow



Note. Adapted from Larsson, 2021

According to Larsson (2021), the development of ethical guidelines for the use of AI can be seen in two ways. The first is that it could be a response to the rapid growth of AI methods. But on the other hand, the adoption of ethical guidelines could be interpreted as an alternative to legislation, making the procedural phase more complicated, which in the long run will require incentives from the supervisory authority to develop techniques for the practical application of existing regulations (cf. Larsson, 2018).

2.4 TRUSTWORTHY AI

Trust plays a fundamental role in user acceptance of AI systems (Barthneck et al., 2021). According to Lee and See (2004), the concept of 'trust' refers to an attitude adopted by one individual towards another in situations of vulnerability. However, in the context of AI, this definition differs, since human interpersonal relationships are moulded by shared experiences, norms and common values. On the other hand, in the case of AI systems, the question arises as

to whether the machine decides autonomously or acts in accordance with previously programmed behaviour (Bartneck et al., 2021).

As mentioned by Lee and See (2004) users adjust their trust in AI systems based on various factors, with reliability being the predominant factor that determines an individual's level of trust or distrust towards AI systems. On the other hand, Hancock et al. (2011), actors that influence users' trust in automated systems include system reliability, false alarm rate and others.

In the context of regulation, trust is also directly linked to the protection of personal data. In the United States, the Patriot Act requires companies to provide the government with access to data stored on cloud servers. This requirement has caused discomfort among European customers of American companies, who have challenged the sharing of personal data with the US government (Barthneck et al., 2021).

As noted by Barthneck et al. (2021), in 2015 the Court of Justice of the European Union annulled the Safe Harbor agreement between the US and the EU, as it did not offer sufficient guarantees to protect the personal data of European citizens. In response to this decision and with the aim of regaining the trust of their customers, several cloud storage companies opened data centres in Europe, ensuring that their European customers' data would be managed in accordance with the European standards implemented.

2.4.1 How to Implement Trustworthy AI

Numerous international organizations and regulatory bodies have published ethical guidelines and codes of conduct intended to foster public trust in artificial intelligence. However, as noted by Kerasidou et al. (2022), many of these initiatives focus more on cultivating trust than on ensuring that AI systems are inherently trustworthy. This distinction is crucial, as it suggests that promoting public confidence does not necessarily imply that ethical standards are being met.

The White Paper by the European Commission on “On Artificial Intelligence: A European Approach to Excellence and Trust” defines the “Ecosystem of trust”. This approach seeks to provide citizens with the confidence to adopt AI applications while offering companies and public organizations the legal clarity to innovate responsibility. (European Commission, 2020).

An appropriate, predictable, and enforceable regulatory structure is essential for building trust among citizens and ensuring legal certainty for companies. Such alignment between ethical leadership and innovation can become a competitive advantage for European actors on the global stage (European Commission, 2018). To achieve this, AI systems must be developed in a way that embodies human-like cognitive and ethical capacities such as creativity, judgement, intuition, and responsibility, which contribute to making informed and reasoned decisions (Yun et al., 2016; Lukowicz & Slusallek, 2018).

Nonetheless, the implementation of trustworthy AI is challenged by the opacity ("black box" effect), complexity, and unpredictability of many systems. These characteristics can limit transparency, hinder accountability, and complicate compliance with existing laws protecting fundamental rights (Medium, 2020). In response to such concerns, the European Commission's White Paper outlines seven key requirements for trustworthy AI: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination, and fairness; societal and environmental well-being; and accountability (European Commission, 2020).

In response, the European Commission´s White Paper outlines seven key requirements for trustworthy AI: (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination and fairness; (6) societal and environmental well-being; and (7) accountability (European Commission, 2020).

The Toronto Declaration (Amnesty International & Access Now, 2018) goes further, advising that where transparency cannot be assured, particularly in high risks contexts, like policing or immigration, AI systems should not be deployed. Similarly, UNESCO (2021) recommends ongoing monitoring and accountability throughout that AI system lifecycle to ensure ethical compliance. Municipal governments have also begun to translate these principles into practice. For example, the city of Amsterdam introduced contractual clauses requiring technical transparency, procedural transparency (how algorithms work), and explicability (how decisions affect individuals).

However, trust in government does not necessarily correlate with support for regulation in all domains in areas like transparency and public trust in policymakers has shown little influence on attitudes toward stricter AI oversight (Davidovic & Harring, 2020; Hammar & Jagers, 2006; Tosun et al., 2020; Dietz et al., 2007). Evidence suggests that high trust in tech

companies, while not diminishing support for hard regulation, increases support for soft, market-based instruments—such as labelling and public information. Yet, the less citizens trust these companies, the more skeptical they become about the efficacy of soft regulation (Davidovic & Harring, 2020).

Public outreach initiatives, such as AI literacy campaigns and awareness progams can help rebuild trust by demystifying AI systems and promoting transparency (Djeffal et al., 2022). At the European level, the ethics guidelines developed by the AI HLEG (European Commission, 2019) define trustworthy AI as having three core components: it must be lawful, ethical, and robust.

Ulnicane (2022) explains that these components align with widely accepted values across multiple international guidelines (Jobin et al., 2019). "Lawful" implies compliance with existing legal frameworks; "ethical" requires adherence to human rights and moral standards; and "robust" refers to the technical and social reliability of AI systems, ensuring they do not unintentionally cause harm. Despite these clear principles, the guidelines do not deeply address the issue of legality—an omission that has drawn criticism. However, as noted by the AI HLEG, most AI practices in Europe have already fallen under established legislation, such as the GDPR (2016), the Charter of Fundamental Rights, and various consumer protection laws.

In sum, trustworthy AI depends on the continuous alignment of technological capabilities with legal, ethical, and social standards. Implementation requires not only guidelines but also enforceable safeguards, governance mechanisms, and societal oversight throughout the AI lifecycle.

2.4.2 The Legalization of Ethic Guidelines

The "legalization problem" refers to the tendency within ethical guidelines to adopt a legalistic style or language without offering sufficient ethical justification. Eriksson et al. (2008) identify three main manifestations of this issue:

1. Ethical regulations that are like legal directives but lack argumentative support.
2. Guidelines that only list principles intended for legal codification rather than practical ethical deliberation.
3. Documents that emphasize ethical requirements as prevailing legal ones.

The study of Eriksson et al. (2008) shows some examples of this problem, include the Swedish Act on Ethical Review, which imposes rules on consent and data use without clarifying their rationale, and UNESCO declarations that are intended as frameworks for legal enactment rather than purely ethical documents. Additionally, the authors refer that some organizations such as the World Medical Association and the Pontifical Academy, assert that certain ethical obligations, especially those related to human dignity and fundamental rights should take precedence over legal norms. These tensions illustrate that ethical guidelines often struggle to balance normative clarity with legal enforceability, potentially undermining their intended moral authority.

3. METHODOLOGY

3.1 STRUCTURE OF METHODOLOGICAL APPROACH

To guide the literature review, this dissertation adopted the research framework proposed by Templier and Paré (2015), (Figure 2), which offers a structured approach for systematic literature review in Information Systems. This model was selected due to its emphasis on both conceptual clarity and methodological rigor, making it particularly suitable for topics that interest technology, ethics, and regulation, such as AI.

The framework consists of four phases: selecting databases, identifying keywords, establishing key indicators, and conducting thematic and content analysis. This structure was appropriate given the objective of this dissertation to examine how regulation influences trust in AI systems, through the lens of ethical principles.

Figure 2 - Research Framework

1. Select database Web of Science (WOS)	**2. Collect data (keywords)** AI OR Artificial Intelligence Ethic Regulation Transparency Policy Accountability Trust
3. Indicatives Data analysis on the principal keywords, categories, language, type of documents, countries	**4. Analysis** Discussion of topics Content Analysis Identify research gaps Potential research directions

Note. Content based on Templier, Mathieu and Paré, Guy, 2015

To ensure the reliability of this review, the WOS database selected as the primary source of data collection. This choice was based on its extensive indexing of high-quality, peer-reviewed publications across multiple disciplines.

The search strategy was based on a combination of keywords: AI" OR "Artificial Intelligence", "Ethic" and "Regulation", "Transparency", "policy", "accountability" and "trust". Boolean operators (e.g., OR) were used to expand the scope of the search.

A structured filtering process was applied to refine the initial results (Table 2). The main criteria included:

- Timeframe: 1960 to 2022.
- Document type: Articles and review articles.
- Languages: Portuguese and English.
- Subject categories: Studies directly related to the research topic.
- Geographic scope: Publications from selected countries (Japan, USA, Portugal, Spain, Russia, England, China).

The process resulted in a progressive reduction of results:

- Initial search results: 14.961.
- After applying all filters: 959 relevant articles.

To ensure relevance, articles were manually screened to exclude those that did not directly address the topic. Fields considered unrelated to core focus, such as environmental sciences, optic, and mathematics were excluded. Similarly, bocks and bock chapters were not considered at this stage. A summary of this filtering is presented in Table 2.

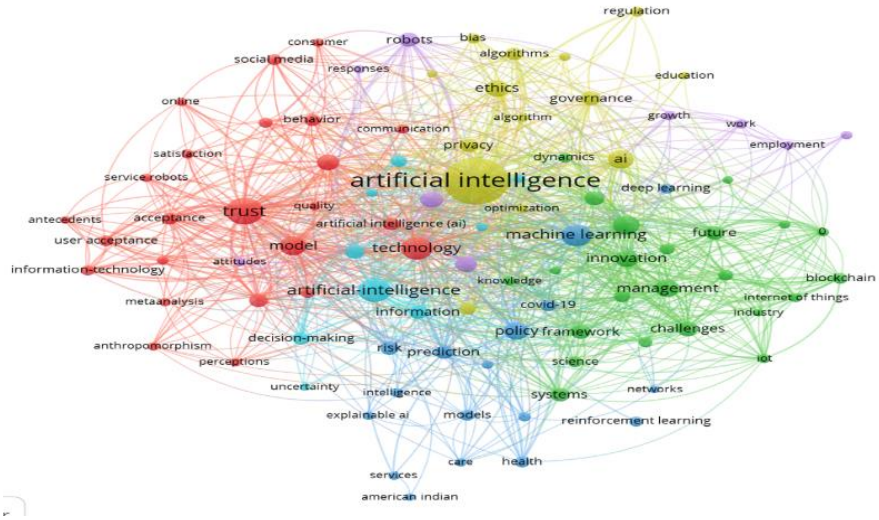
Filters		Number of Publications
Keywords	(“Artificial Intelligence”, “Ethic” and “Regulation”, “Transparency”, “policy”, “accountability” and “trust”)	14961
Time	1960-2022	10093
Type of documents	Article and Review Article	7478
Language	Portuguese and English	7198
Categories	Environmental Sciences, Management, Business, Computer Science Theory Methods, Economics, Social Sciences Interdisciplinary, Ethics, Political Science, Social Issues, Philosophy, Public Administration, International Relations, Health Policy Services, Psychology Experimental, Sociology, Medical Ethics, Social Works, Statistics Probability, Optics, Social Sciences Mathematical Methods, Psychology Social, Education Scientific Disciplines, Development Studies, History, Culture Studies, Demography, Linguistics	1764
Countries	Japan, USA, Portugal, Spain, Russian, England, Peoples R China	959

Table 2 - Results of the filtering process

The visual map generated using the software VOS Viewer illustrates the relationships between the most frequently the relationships between the most frequently occurring keywords in the 959 selected publications. (Figure 3) In this map, each node represents a keyword, while the proximity of the connecting lines indicates the strength of occurrence. Clusters with different colors represent distinct thematic areas.

This visualization provides an overview of how ethical and regulatory concerns are addressed in AI, helping to identify key areas of focus in the current literature.

Figure 3 - Keywords extracted from VOS Viewer



3.2 Comparative Analysis of Ethical Guidelines

To evaluate the relevance and practical application of ethical AI guidelines, this study compares two sources:

- Jobin et al. (2019): Identified five core ethical principles by analyzing 84 public and private AI ethic initiatives.
- Rothenverger & Fabian (2019): Applied a mixed-methods approach to assess the perceived importance of various AI ethical principles, using qualitative interviews and a quantitative citizen survey (n=51).

Rothenberg and Benjamim Fabian (2019) followed the Grounded Theory Methodology (GTM). In the first phase, seven qualitative interviews were conducted with internal and external experts from the ICT (Information and Communication Technology) sector. Following the exploratory phase, a quantitative questionnaire was conducted with 51 citizens. The

relevance of each guideline was calculated using a weighted arithmetic mean combining both groups:

$$\emptyset = (\emptyset Experts \times 0{,}5) + (\emptyset Survey \times 0{,}5)$$

The ethical guidelines assessed in the study were compiled from academic, industrial, governmental, and associative sources. The table 3 summarizes the six key principles identified, their definitions, and the type of organizations that proposed them.

Table 3 - Selected Ethical Guidelines

Guideline	**Type of Organization**	**Definition**
Transparency	Government, Association, Research and Development institution	An AI system must be transparent about being an AI (before usage or interaction). Therefore, an international standard must be launched.
Responsibility	Association	The operator and user of an AI are bearing the blame for any action (and their consequences) of the AI system.
Protection of Data Privacy	Government, Industry	Unauthorized interceptions should be avoided. The user must agree explicitly to the usage of his private data.
Bias should be minimized	Industry	Unfair, racist bias should be minimized.
An Ai should have a purpose	Industry, Association	Supporting the human should be the highest purpose of an AI. The AI must not replace the human. A human-machine cooperation model will be established.
Robustness	Academic, Association, Industry	AI algorithms should be robust against manipulations, both internally and externally. For example, a language assistant should not order any items through any external influences.

Note. Adapted from Lea Rothenberg, Benjamin Fabian & Elmar Arunov, 2019

To align with the structure of Jobin et al. (2019), the principals “Justice, Fairness and Equity” and “Non-Maleficence” were grouped under the category “Bias Should Be Minimized”, due to conceptual affinity with Rothenberher and Fabian´s definitions.

These six guidelines were then evaluated by the interviewed experts, who were asked to rate the importance. The table 4 represents the ranking of the guidelines based on the arithmetic mean of the expert responses.

Table 4 - Ranking of Ethical Guidelines by the Experts

Guidelines	Arithmetic Mean	Rank Order
Responsibility	4.71	1
Transparency	4.43	2
Protection of Data Privacy	4.43	2
Robustness	4.14	3
Bias should be minimized	3.67	4
An AI should have a purpose	3.29	5

Note. Adapted from Lea Rothenberg, Benjamin Fabian and Elmar Arunov, 2019

Following the expert evaluation, a comparative analysis was carried out using the data from Jobin et. al. (2019), which identified the frequency with ethical principles appeared across 84 public and private AI initiatives. The arithmetic mean for each principle was calculated as follows:

$$\emptyset tramsparency\ (Jobin) = \frac{Number\ of\ documents\ mentioning\ Transparency}{Tota\ of\ Documents} \times 100$$

$$= \frac{73}{84} \times 100 = 86.9\%$$

$$\emptyset Responsability\ (Jobin) = \frac{Number\ of\ documents\ mentioning\ Responsability}{Total\ of\ Documents} \times 100$$

$$= \frac{60}{84} \times 100 = 71.4\%$$

$$\emptyset Privacy\ (Jobin) = \frac{Number\ of\ documents\ mentioning\ Privacy}{Total\ of\ Documents} \times 100 = \frac{47}{84} \times 100 = 55.9\%$$

$$\emptyset Justice, fairness\ and\ equity\ (Jobin)$$

$$= \frac{Number\ of\ documents\ mentioning\ justice, fairness\ and\ euiqty}{Total\ of\ Documents} \times 100$$

$$= \frac{68}{84} \times 100 = 81.0\%$$

$$\emptyset Non - maleficence\ (Jobin)$$
$$= \frac{Number\ of\ documents\ mentioning\ Non - malificience}{Total\ of\ Documents} \times 100 = \frac{60}{84} \times 100$$
$$= 71.4\%$$

Given conceptual overlap between the categories "Justice, Fairness, and Equity" and "Non-Maleficence" in Jobin et. al.´s framework and the category "Bias Should Be Minimized" in Rothenberg & Fabian´s classification, a combined mean was calculated:

$$\emptyset Bias\ should\ be\ minimized\ (Jobin) = \frac{81.4 + 71.4}{4} = 76.2\%$$

Based on these percentages, scores were rescaled to match Rothenberger & Fabian´s 5-point scale, resulting in the table below:

Table 5 - Rescaled Scores and Ranking of Ethical Guidelines Based on Expert Evaluation

Guideline	Arithmetic Mean (%)	Arithmetic Mean (%)	Rank order
Transparency	86.9%	4.34	1
Bias should be minimized	76.2%	3.81	2
Responsibility	71.4%	3.57	3
Privacy	55.9%	2.79	4

The principles "Robustness" and "An AI Should Have a Purpose" were excluded from the comparative phase, as they were not featured in the Jobin et al. (2019) dataset.

To align both data sources, a combined final score was calculated using the same weighted average formula as in Rothenberger & Fabian´s study:

$$\emptyset final = (\emptyset Jobin \times w1) + (\emptyset Experts \times w2)$$

The resulting rankings are shown below:

Table 6 - Combined Scores and Ranking of Ethical Guidelines (Jobin and Expert Evaluation)

Guidelines	Arithmetic Mean	Rank Order
Transparency	4.38	1

Responsibility	4.14	2
Bias should be minimized	3,70	3
Privacy	3.61	4

This integrative analysis provides an empirical basis for identifying which ethical principles are perceived as most relevant, both in expert evaluations in relation to regulatory alignment.

To address the research question of the present dissertation a third dimension was introduced: the degree of alignment between each principle and existing legal or regulatory frameworks. Relevant regulatory instruments up to 2022 were reviewed to assess this alignment, including the following:

- The GDPR (2016), establishes the rights to privacy, data protection, and transparency.
- The White Paper (2020), outlines an "ecosystem of trust" and key ethical requirements for trustworthy AI.
- The Toronto Declaration (2018), emphasizes transparency, accountability, and human rights protections in the context of algorithmic decision-making.

Each principle was assigned a regulatory alignment score ranging from 0 to 1:

- 0.0 = no clear regulatory support identified by 2022.
- 0.5 = partially covered through general principles.
- 1.0 = clearly codified in binding legislation.

The regulatory scores were defined as follow:

Table 7 - Regulatory score

Guideline	**Regulation Score**	**Justification**
Transparency	1.0	Addressed in GDPR (2016), (e.g., "right to explanation") and emphasized in both White Paper (2020) and the Toronto Declaration (2018)
Responsibility	1.0	Addressed in the White Paper (2020) as a key requirement for trustworthy AI and the GDPR (2016)
Privacy	1.0	Addressed in GDPR (2016), which legally protects personal data and user consent, and reinforced in the White Paper´s (2020) ethical framework

Bias Should be Minimized	0.5	Addressed in the Toronto Declaration (2018) through anti-discrimination principles and the fairness discourse in the White Paper (2020)

The values were then integrated into a final composite score for each principle, using the following weighted formula:

$$\emptyset total = (\emptyset Jobin \times 0.33) + (\emptyset Experts \times 0.33) + (\emptyset Regulation \times 0.33)$$

The resulting score are presented in the table below:

Table 8 - Final Ranking of Ethical Guidelines

Guideline	**Jobin Score**	**Expert Score**	**Regulation Score**	**Final Score**	**Rank order**
Transparency	4.34	4.43	1.0	3.22	1
Responsibility	3.57	4.71	1.0	3.06	2
Bias Should be Minimized	3.81	3.67	0.5	2.71	3
Privacy	2.79	4.43	1.0	2.63	4

The purpose of this comparative framework to evaluate the perceived relevance of ethical principles in AI governance, but also analyze how alignment with binding legal instruments contributes to their credibility and potential to promote trust.

4. RESULTS

This chapter presents the empirical results of the comparative analysis of ethical AI guidelines, developed through a triangulated methodology that combines academic literature, expert evaluation, and regulatory alignment. The objective is to identify which ethical principles are most frequently referenced, most valued by experts, and most supported by legal instruments in force until 2022. The analysis is based on two key sources:

(1) Jobin et al. (2019) - who analyzed 84 public and private AI initiatives to identify ethical principles in AI governance documents.
(2) Rothenberger and Fabian (2019) - who conducted a mixed-methods study combining qualitative interviews with experts and a quantitative citizens survey to evaluate the importance of various ethical principles.

To enhance the robustness of the analysis, a third variable was added:

(3) Regulation alignment of each principle with existing regulatory instruments, such as the GDPR (2016), the White Paper (2020), and the Toronto Declaration (2018).

To generate a comprehensive and comparative metric, a composite score was calculated based on a weighted average of the following dimensions, the frequency of appearance in literature (Jobin et al. 2019), expert prioritization (Rothenberger & Fabian, 2019) and legal support through regulation (Regulation Score). The following sections present the data and results of this composite analysis.

4.1 DATA PRESENTATION

This section presents the main results of the comparative analysis between ethical AI guidelines, expert perspectives and relevant regulation.

4.1.1 FREQUENCY IN THE LITERATURE (JOBIN ET AL., 2019)

The table below presents the percentage of occurrence of the core ethical principles in the 84 initiatives analyzed by Jobin et al. (2019), along with their corresponding values on a 5-point scale:

Table 9 – Frequency in institutional documents (Jobin et. al., 2019)

Guideline	Occurrence (%)	Rescaled Mean (0-5)
Transparency	86.9%	4.34
Responsibility	71.2%	3.81
Bias Should Be Minimized	76.2% (combined)	3.57
Privacy	55.9%	2.79

4.1.2 Expert Evaluation (Rothenberg & Fabian, 2019)

The following table summarizes how each ethical principle was ranked by experts through qualitative interviews and a quantitative survey:

Table 10 – Expert evaluation (Rothenberg & Fabian, 2019)

Guideline	Expert Mean (0-5)	Rank order
Responsibility	4.71	1
Transparency	4.43	2
Privacy	4.43	3
Bias Should Be Minimized	3.67	4

4.1.3 Combined Score (Literature + Experts)

A combined score was calculated for each principle using a weighted arithmetic mean:

$$\emptyset combined = (\emptyset Jobin \times 0.5) + (\emptyset Experts \times 0.5)$$

Table 11 – Combined score (Literature + Experts)

Guideline	Jobin score	Expert score	Combined Mean	Rank order
Transparency	4.34	4.43	4.38	1
Responsibility	3.57	4.71	4.14	2
Bias Should Be Minimized	3.81	3.67	3.74	3
Privacy	2.79	4.43	3.61	4

4.1.4 Final Score Including Regulatory Alignment

To assess institutional applicability, each principle was assigned a regulation score (0.0 = no clear regulatory support; 0.5 = partially covered thought general principles; 1.0 = clearly codified in binding legislation). These were integrated as follow:

$$\emptyset final = (\emptyset Jobin \times 0.33) + (\emptyset Experts \times 0.33) + (\emptyset Regulation \times 0.33)$$

Table 12 – Final score including regulatory alignment

Guideline	Jobin	Expert	Regulation	Final Mean	Rank order
Transparency	4.34	4.43	1.00	4.54	1
Responsibility	3.57	4.71	1.00	4.38	2
Bias Should Be Minimized	3.81	3.67	0.5	3.29	3
Privacy	2.79	4.43	0.5	3.21	4

4.2 Discussion on the Results

The results reveal important insights regarding the prioritization of ethical principles in the literature, the expert opinion, and regulation, in the field of AI.

First, transparency emerges consistently as the most prominent principle across all dimensions analyzed. It ranks highest in frequency within institutional guidelines (Jobin et al., 2019), is strongly supported by expert evaluation (Rothenberger & Fabian, 2019), and is the most embedded principle in legal frameworks, such as GDPR (2016). This suggest that transparency is not only widely endorsed but also actively prioritized by regulators, possible because of its foundational role in enabling trustworthy AI systems, especially when transparency mechanisms allow users and regulators to scrutinize decision making processes.

In second place is responsibility, which, although less frequent in the literature, receives the highest score from experts and strong regulatory support. This may reflect a growing concern with accountability mechanisms, such as assigning responsibility for harms or errors in AI use.

Bias Should Be Minimized ranks third, despite its relatively high frequency in literature. Its low regulatory score indicates that legal systems are still catching up in translating fairness and non-discrimination principles into enforceable norms. While bias mitigation is widely

recognized as essential, especially in high-risk contexts, its limited legal anchoring may reduce its impact on institutional trust.

Privacy places the fourth position. Although strongly supported by both experts and regulation, appears less prominently in ethical guidelines. This misalignment may reflect an assumption that privacy is already covered by general data protection legislation and does not require repeated ethical emphasis. However, this might lead to a disconnection between regulatory priorities and ethical discourses, potentially affecting how consistently the principle is applied in AI development.

Overall, the findings reveal a partial convergence between literature, expert perception, and regulatory frameworks, particularly around transparency and responsibility. More importantly, the analysis may suggest that legal frameworks enhance not only the visibility but also the credibility of ethical principles, increasing their potential to build trust in AI.

5. CONCLUSIONS AND FUTURE RESEARCH

This dissertation examined the convergence between ethical guidelines for AI and existing regulatory framework, with a particular focus on the role of regulation as a potential enabler of trust. Grounded in the comparative analysis of two key studies, Jobin et. Al (2019) and Rothenberg & Fabian (2019), this research required us to identify the most recurrent ethical principles, assess their perceived relevance among experts and citizens, and evaluate their alignment with existing legislation.

In response to the research question "Is regulation a value of trust?" the findings suggest that regulation can indeed function as a foundational value in fostering trust in AI. This is particularly evident when regulation embodies key principles such as transparency, responsibility, and fairness, and is implemented clearly, predictable, and ethically grounded.

The analysis revealed that transparency and responsibility are consistently prioritized across both ethical discourse and legal anchoring, indicating a strong convergence between normative aspirations and regulatory intent. However, some ethical principles, such as bias minimization, remain underrepresented in formal legislation. This reveals a gap between their theoretical importance and legal enforceability, highlighting the challenge of translating ethical intentions into concrete, binding rules.

The central takeaway of this study is that regulation does not merely act as a legal constraint but can serve as a normative value that legitimizes and reinforces trust in AI. By embedding ethical expectations into law, regulation reduces uncertainty and promotes public confidence in AI systems.

Despite these contributions, this research has several limitations. First, it is based on secondary data and comparative analysis, without direct empirical validation through interviews or fieldwork. Second, although the study is based on widely cited ethical frameworks, it does not examine how these principles are interpreted and applied within specific sectors, such as healthcare, finance, or public administration, where the ethical implications of AI differ significantly. Third, the analysis excludes relevant perspectives from different linguistic or cultural contexts.

Future research should address these gaps by incorporating empirical studies into how ethical principles are implemented in practice, particularly in high risks domains. It would also be valuable to explore how national regulatory strategies align with the recent European legislative, such as the EU AI Act, to promote effectiveness in AI governance.

Finally, while ethical guidelines are essential in shaping the values that guide AI development, their impact depends on how effectively they are integrated into enforceable regulatory systems. Regulation, when designed with ethical intent, has the potential not only to guide AI behavior, but also to embody good values, such as trust.

6. BIBLIOGRAPHICAL REFERENCES

Abrassart, C., Bengio, Y., Chicoisne, G., de Marcellis-Warin, N., Dilhac, M. A., Gambs, S., Gautrais, V., Gibert, M., Langlois, L., Laviolette, F., Lehoux, P., Maclure, J., Martel, M., Pineau, J., Railton, P., Régis, C., Tappolet, C., and Voarino, N. (2018). Montreal Declaration for a Responsible Development of Artificial Intelligence. https://montrealdeclaration-responsibleai.com/the-declaration/

Amnesty International & Access Now. (2018). The Toronto Declaration: Protecting the rights to equality and non-discrimination in machine learning systems. https://www.amnesty.org/en/documents/pol30/8447/2018/en/

Braun, T., Fung, B. C., Iqbal, F., & Shah, B. (2018). Security and privacy challenges in smart cities. Sustainable Cities and Society, 39, 499–507. https://doi.org/10.1016/j.scs.2018.02.039

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Conference on Fairness, Accountability and Transparency, PMLR (Proceedings of Machine Learning Research, 81, 77–91). https://proceedings.mlr.press/v81/buolamwini18a.html

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356 (6334), 183-186. https://doi.org/10.1126/science.aal4230

Chiao, V. (2019). Fairness, accountability and transparency: notes on algorithmic decision-making in criminal justice. International Journal of Law in Context, 15(2), 126-139. http://doi.org/10.1017/S1744552319000077.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), 153-163 https://doi.org/10.1089/big.2016.0047

Coeckelbergh, M. (2019). Artificial intelligence: Some ethical issues and regulatory challenges. Technology and regulation, 2019, 31-34. https://doi.org/10.71265/a9yxhg88

Council of Europe. (2017). Guidelines on the protection of individuals with regard to the processing of personal data in a world of Big Data. https://rm.coe.int/16806ebe7a

Davidovic, D., & Harring, N. (2020). Exploring the cross-national variation in public support for climate policies in Europe: The role of quality of government and trust. Energy Research & Social Science, 70, 101785. https://doi.org/10.1016/j.erss.2020.101785

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(3): 505–25. https://doi.org/10.1007/s10676-021-09593-z

Djeffal, C., Siewert, M. B., & Wurster, S. (2022). Role of the state and responsibility in governing artificial intelligence: a comparative analysis of AI strategies. Journal of European Public Policy, 29(11), 1799-1821. https://10.1080/13501763.2022.2094987

Domingos, P. (2015). The master algorithm: How the quest for the ultimate learning machine will remake our world. Basic Books. https://www.redalyc.org/articulo.oa?id=638067264018

Du, S., & Xie, C. (2021). Paradoxes of artificial intelligence in consumer markets: Ethical challenges and opportunities. Journal of Business Research, 129, 961-974. https://doi.org/10.1016/j.jbusres.2020.08.024

Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a'right to an explanation'is probably not the remedy you are looking for. Duke L. & Tech. Rev., 16, 18. https://ssrn.com/abstract=2972855

Eriksson, S., Höglund, A. T., & Helgesson, G. (2008). Do ethical guidelines give guidance? A critical examination of eight ethics regulations. Cambridge Quarterly of Healthcare Ethics, 17(1), 15-29. https://www.researchgate.net/profile/Stefan-Eriksson-2/publication/5388143_Do_Ethical_Guidelines_Give_Guidance_A_Critical_Examination_of_Eight_Ethics_Regulations/links/0c96052a731665f8f7000000/Do-Ethical-Guidelines-Give-Guidance-A-Critical-Examination-of-Eight-Ethics-Regulations.pdf

European Commission. (2020). White paper on artificial intelligence: A European approach to excellence and trust. (COM (2020) 65 final) https://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

European Commission: Directorate-General for Communications Networks, Content and Technology. (2019). Ethics guidelines for trustworthy AI. Publications Office. https://data.europa.eu/doi/10.2759/346720

European Commission. (2019). Ethics Guidelines for Trustworthy AI. High-Level Expert Group on Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Commission. (2018). Artificial Intelligence for Europe. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=COM%3A2018%3A237%3AFIN

Eyert, F., Irgmaier, F., & Ulbricht, L. (2022). Extending the framework of algorithmic regulation. The Uber case. Regulation & Governance, 16(1), 23-44. https://doi.org/10.1111/rego.12371

Falco, G., Viswanathan, A., Caldera, C., & Shrobe, H. (2018). A master attack methodology for an AI-based automated attack planner for smart cities. IEEE Access, 6, 48360-48373. https://doi.org/10.1109/ACCESS.2018.2867556

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 (Berkman Klein Center Research Publication No. 2020-1). Berkman Klein Center for Internet & Society. https://doi.org/10.2139/ssrn.3518482

Gibbs, S. (2017). Elon Musk: regulate AI to combat ‘existential threat’ before it’s too late. The Guardian. https://www.theguardian.com/technology/2017/jul/17/elon-musk-regulation-ai-combat-existential-threat-tesla-spacex-ceo

Gupta, A., Mishra, M. (2022). Ethical Concerns While Using Artificial Intelligence in Recruitment of Employees. Business Ethics and Leadership, 6(2), 6-11. http://doi.org/10.21272/bel.6(2).6-11.2022

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. Minds and machines, 30(1), 99-120. https://doi.org/10.1007/s11023-020-09517-8

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. Human factors, 53(5), 517-527. https://doi.org/10.1177/0018720811417254

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

Joshua A. Kroll , Joanna Huey , Solon Barocas , Edward W. Felten , Joel R. Reidenberg , David G. Robinson & Harlan Yu Accountable Algorithms, 165 U. Pa. L. Rev. 633 (2017).
https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3

Kazim, E., & Koshiyama, A. S. (2021). A high-level overview of AI ethics. Patterns, 2(9). https://doi.org/10.1016/j.patter.2021.100314

Kerasidou, C. X., Kerasidou, A., Büscher, M., & Wilkinson, S. (2022). Before and beyond trust: reliance in medical AI. Journal of Medical Ethics, 48(11), 852–856. https://doi.org/10.1136/medethics-2020-107095

Khalifa, E. (2019). Smart cities: Opportunities, challenges, and security threats. Journal of Strategic Innovation and Sustainability, 14, 79–88. https://doi.org/10.33423/jsis.v14i3.2108

Kokolakis, S. (2017). Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. Computers & security, 64, 122-134. http://dx.doi.org/10.1016/j.cose.2015.07.002

Koulu, R. (2020). Human control over automation: EU Policy and AI Ethics. European Journal of Legal Studies, 12(1), 9-46. https://doi.org/10.2924/EJLS.2019.019

Larsson, S. (2021). AI in the EU: Ethical Guidelines as a Governance Tool. In A. Bakardjieva Engelbrekt, K. Leijon, A. Michalski, & L. Oxelheim (Eds.), The European Union and the Technology Shift (pp. 85-110). (Palgrave Macmillan). Springer Nature. https://doi.org/10.1007/978-3-030-63672-2_4

Larsson, S. (2020). On the governance of artificial intelligence through ethics guidelines. Asian Journal of Law and Society, 7(3), 437–451. https://doi.org/10.1017/als.2020.19

Larsson, S. (2018). Algorithmic governance and the need for consumer empowerment in data-driven markets. Internet Policy Review, 7(2). https://doi.org/10.14763/2018.2.791

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46 (1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Lu, Y., & Da Xu, L. (2018). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103-2115. https://doi.org/10.1109/JIOT.2018.2869847

Lukowicz, P., & Slusallek, P. (2018). How to avoid an AI interaction singularity. Interactions, 25 (5), 72-78. https://doi.org/10.1145/3264995

Macdonald, D. (2019). Why low trust in government may mean Americans don't want anything done about inequality. USApp-American Politics and Policy Blog. London School of Economics.https://eprints.lse.ac.uk/103292/1/usappblog_2019_12_23_why_low_trust_in_government_may_mean_americans.pdf

Matus, K. J., & Veale, M. (2022). Certification systems for machine learning: Lessons from sustainability. Regulation & Governance, 16(1), 177-196. https://doi.org/doi:10.1111/rego.12417

McCarthy, J. (2007). What is Artificial Intelligence? Stanford University. http://www-formal.stanford.edu/jmc/

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 2053951716679679. https://doi.org/10.1177/2053951716679679

Morishita, L., & van Zyl, D. (2017). Exploring the significance of earning a social license to operate in an urban setting. In Proceedings of the 8th International Conference on Sustainable Development in the Minerals Industry. https://doi.org/10.15273/gree.2017.02.048

Morley, J., Machado, C. C., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: a mapping review. Social science & medicine, 260, 113172. https://doi.org/10.1016/j.socscimed.2020.113172

Olhede, S. C., & Wolfe, P. J. (2018). The growing ubiquity of algorithms in society: implications, impacts and innovations. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376(2128), 20170364. http://dx.doi.org/10.1098/rsta.2017.0364

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Publishing Group. https://dl.acm.org/doi/10.5555/3002861

Pandey, A., & Caliskan, A. (2021). Disparate impact of artificial intelligence bias in ridehailing economy's price discrimination algorithms. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (pp. 822-833). https://doi.org/10.1145/3461702.3462561

Pasquale, F. (2015). The Black Box Society, the Secret Algorithms That Control Money and Information. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Pauwels, E. (2020). Artificial intelligence and data capture technologies in violence and conflict prevention. Global Centre on Cooperative Security, 10, 2022. https://globalcenter.org/wp-content/uploads/GCCS_AIData_PB_H-1.pdf

Prunkl, C., & Whittlestone, J. (2020, February). Beyond near-and long-term: Towards a clearer account of research priorities in AI ethics and society. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 138-143). https://doi.org/10.1145/3375627.3375803

Razmi, T. (2020). Summary of EU white paper on artificial intelligence: A European approach to excellence and trust. Medium. https://medium.com/@tibastar/summary-of-eu-white-paper-on-artificial-intelligence-a-european-approach-to-excellence-and-trust-e04a1a018b5

Rothenberger, Lea; Fabian, Benjamin; and Arunov, Elmar, (2019). "Relavance of Ethical Guidelines for Artificial Intelligence – A survey and evaluation". In Proceedings of the 27th European Conference on Information Systems (ECIS), Stockholm & Uppsala, Sweden, June 8-14, 2019. ISBN 978-1-7336325-0-8 Research-in-Progress Papers. https://aisel.aisnet.org/ecis2019_rip/26

Saghiri, A. M., Vahidipour, S. M., Jabbarpour, M. R., Sookhak, M., & Forestiero, A. (2022). A survey of artificial intelligence challenges: Analyzing the definitions, relationships, and evolutions. Applied sciences, 12(8), 4054. https://doi.org/10.3390/app12084054

Salmon, P. M., Hancock, P., & Carden, A. W. (2019). To protect us from the risks of advanced artificial intelligence, we need to act now. The Conversation. https://theconversation.com/to-protect-us-from-the-risks-of-advanced-artificial-intelligence-we-need-to-act-now-107615

Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2020). What's next for AI ethics, policy, and governance? A global overview. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), 153–158. https://doi.org/10.1145/3375627.3375804

Schiff, D., Borenstein, J., Biddle, J., & Laas, K. (2020). What's next for AI ethics, policy, and governance? A global overview. In Proceedings of the AIES '20: AAAI/ACM Conference on AI, Ethics, and Society

Schleiger, E., & Hajkowicz, S. (2019). Artificial intelligence in Australia needs to get ethical, so we have a plan. The Mandarin. https://www.themandarin.com.au/107060-artificial-intelligence-in-australia-needs-to-get-ethical-so-we-have-a-plan/

Selbst, A., & Powles, J. (2018). "Meaningful Information" and Rights to Explanation. Proceedings of the 1st Conference of Fairness, Accountability and Transparency, PMLR 81:48-48

Sweeney, L. (2013). Discrimination in online ad delivery. Communications of the ACM, 56(5), 44-54.https://doi.org/10.1145/2447976.2447990

Templier, M., & Paré, G. (2015). A framework for guiding and evaluating literature reviews. Communications of the Association for Information Systems, 37(1), 6. https://doi.org/10.17705/1CAIS.03706

UK Parliament. (n.d.). House of Lords. https://www.parliament.uk/business/lords/

Ulnicane, I. (2022). Artificial intelligence in the European Union: Policy, ethics and regulation. In T. Hoerber, G. Weber, & I. Cabras (Eds.), The Routledge handbook of European integrations (Chapter 14). Routledge. https://doi.org/10.4324/9780429262081-19

Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., & Wanjiku, W.-G. (2021). Framing governance for a contested emerging technology: Insights from AI policy. Policy and Society, 40 (2), 158 – 177. https://doi.org/10.1080/14494035.2020.1855800

UNESCO. (2021). Report of the Social and Human Sciences Commission (SHS): 41 C/73. UNESCO General Conference, 41st session. https://unesdoc.unesco.org/ark:/48223/pf0000379920

Valle-Cruz, D., Fernandez-Cortez, V., & Gil-Garcia, J. R. (2022). From E-budgeting to smart budgeting: Exploring the potential of artificial intelligence in government decision-making for resource allocation. Government Information Quarterly, 39 (2), 101644. https://doi.org/10.1016/j.giq.2021.101644

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. Big Data & Society, 4 (2), 1 –17. https://doi.org/10.1177/2053951717743530

Villani, C., Bonnet, Y., Schoenauer, M., Berthet, C., Cornut, A.-C., Levin, F., & Rondepierre, B. (2018). For a meaningful artificial intelligence: Towards a French and European strategy. Conseil national du numérique. https://www.jaist.ac.jp/~bao/AI/OtherAIstrategies/MissionVillani_Report_ENG-VF.pdf

Weng, S., Schwarz, G., Schwarz, S., & Hardy, B. (2021). A Framework for Government Response to Social Media Participation in Public Policy Making: Evidence from China. International Journal of Public Administration, 44(16), 1424–1434. https://doi.org/10.1080/01900692.2020.1852569

Willems, J., Schmid, M. J., Vanderelst, D., Vogel, D., & Ebinger, F. (2022). AI-driven public services and the privacy paradox: do citizens really care about their privacy? Public Management Review, 25(11), 2116–2134. https://doi.org/10.1080/14719037.2022.2063934

Wirtz, B. W., Weyerer, J. C., & Kehl, I. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. Government information quarterly, 39(4), 101685. https://doi.org/10.1016/j.giq.2022.101685

Yigitcanlar, T., Desouza, K. C., Butler, L., & Roozkhosh, F. (2020). Contributions and Risks of Artificial Intelligence (AI) in Building Smarter Cities: Insights from a Systematic Review of the Literature. Energies, 13 (6), 1473. https://doi.org/10.3390/en13061473

Yoo, C. S., & Lai, A. (2020). Regulation of algorithmic tools in the United States. Journal of Law & Economic Regulation. https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=3249&context=faculty_scholarship

Yun, J. J., Lee, D., Ahn, H., Park, K., & Yigitcanlar, T. (2016). Not deep learning but autonomous learning of open innovation for sustainable artificial intelligence. Sustainability, 8(8), 797. https://doi.org/10.3390/su8080797

Zuiderveen Borgesius, F. J. (2020). Strengthening legal protection against discrimination by algorithms and artificial intelligence. The International Journal of Human Rights, 24(10), 1572-1593. https://doi.org/10.1080/13642987.2020.1743976

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa