AI and Ethics
https://doi.org/10.1007/s43681-024-00480-z

ORIGINAL RESEARCH

A semi-automated software model to support AI ethics compliance assessment of an AI system guided by ethical principles of AI

Maria Assunta Cappelli[1] · **Giovanna Di Marzo Serugendo**[1]

Received: 15 November 2023 / Accepted: 9 April 2024

Abstract
Compliance with principles and guidelines for ethical AI has a significant impact on companies engaged in the development of artificial intelligence (AI) systems. Specifically, ethics is a broad concept that continuously evolves over time and across cultural and geographical boundaries. International organisations (IOs), individual states, and private groups, all have an interest in defining the concept of ethics of AI. IOs, as well as regional and national bodies, have issued many decisions on AI ethics. Developing a system that complies with the ethical framework poses a complex challenge for companies, and the consequences of not complying with ethical principles can have severe consequences, making compliance with these requirements a key issue for companies. Furthermore, there is a shortage of technical tools to ensure that such AI systems comply with ethical criteria. The scarcity of ethics compliance checking tools for AI, and the current focus on defining ethical guidelines for AI development, has led us to undertake a proposal consisting in a semi-automated software model to verify the ethical compliance of an AI system's code. To implement this model, we focus on the following important aspects: (1) a literature review to identify existing ethical compliance systems, (2) a review of principles and guidelines for ethical AI to determine the international and European views regarding AI ethics, and (3) the identification of commonly accepted principles and sub-principles of AI. These elements served to inform (4) our proposal for the design of a semi-automated software for verifying the ethical compliance of AI systems both at design-time (ethics-by-design perspective) and afterwards on the resulting software.

Keywords AI ethics compliance · AI ethics principles · Ethical conformity · AI ethics regulations

1 Introduction

Compliance with the principles and guidelines for ethical artificial intelligence (AI) significantly impacts companies involved in AI system development. Ethics, being a broad concept, continuously evolves over time and across cultural and geographical boundaries.

Developing a system that complies with these frameworks poses a complex challenge for companies. Non-compliance with ethical and legal requirements can lead to severe consequences, making it a key issue for business.

De Laat et al. [14] found that most US and European companies ruled out government regulation, such as standard-setting, codes of ethics, and legislation. One critical aspect is the use of unbiased data according to ethical criteria. For example, a facial recognition system must be trained on unbiased data to avoid reproducing biases in decisions, violating ethical principles. Therefore, companies need to cleanse the data. Bessens et al. [7] observed that developing an AI system in compliance with ethical principles and legal requirements is a costly process. Startups with close data-sharing partnerships with technology companies were more likely to spend significant resources to remove training data or reject business opportunities to comply with ethical principles.

Converting ethical norms into computational algorithms that are understandable to the system is a recognised challenge due to their complexity and specific nature [30, 39]. To address this challenge, several studies propose using systems to assess AI system's alignment with ethical criteria. While

✉ Maria Assunta Cappelli
maria.cappelli@unige.ch

Giovanna Di Marzo Serugendo
giovanna.dimarzo@unige.ch

[1] Centre Universitaire d'Informatique (CUI), Université de Genève, Rte de Drize 7, 1227 Carouge, Switzerland

Published online: 02 May 2024

manual verification of ethical compliance is often suggested, limited research discusses the adoption of automated systems, which are not yet fully developed. Currently, there are no fully developed automated systems to verify AI's ethical compliance (see Sect. 2).

The workshop entitled "AI informed decision support instrument (AI-DSI) for digital industries" held in Geneva on 22 December 2022, with the participation of companies interested in exploring the ethical compliance of the AI system, revealed a compelling interest in controlling the compliance of AI ethics.

Companies are concerned about assessing the compliance of AI ethics with their systems. Our analysis will focus on key issues. Firstly, ensuring the ethical compliance of AI systems is a challenge. When companies set up an AI system, there is no automatic way to verify compliance, and ethical compliance is often manually checked. The lack of automated tools disincentives the development and commercialisation of AI systems. Secondly, addressing the party assuming the risks is crucial. While ethical guidelines are not mandatory, companies are responsible for creating systems aligned with ethical principles to avoid harm, ensure consumer trust, and ensure respectful and responsible development of AI systems. Therefore, programmers need to identify the ethical risks associated with AI systems and implement measures to mitigate them. Thirdly, considering the ethical principles for implementing an AI system, including explainability, fairness, and accuracy, is essential. Each sector has specific goals and principles, and corresponding ethical principles should be followed. Fourthly, companies will prefer to proactively prepare for ethical compliance as it becomes a duty in the coming years. It's crucial to involve experts in evaluating compliance monitoring systems to ensure ethical decision-making [11]. Lastly, another important issue is the transition from theory to practice, translating high-level ethical principles for AI into applicable software.

The scarcity of ethics compliance checking tools for AI, the current focus on defining ethical guidelines for AI development, and the concerns of some companies in building an ethical AI system led us to undertake a semi-automated software model to verify AI system code's ethical compliance. A semi-automated software is suitable for evaluating ethical compliance, as this process requires subjective value assessments. A software program cannot entirely replace a human being but requires human intervention to determine whether what has been judged as ethically acceptable or unacceptable.

Our approach to address this issue adopted the following steps: (1) a literature review to identify existing ethical compliance systems, (2) a review of principles and guidelines for ethical AI to determine the international and European views regarding AI ethics, and (3) the identification of commonly accepted principles and sub-principles of AI (as shown in Fig. 1). These elements served to inform (4) our proposal for the design of a semi-automated software for verifying the ethical compliance of AI systems.

This paper is structured as follows. Section 2 presents the research conducted to establish an ethical control system for an AI system. Additionally, a sub-section includes studies carried out for general compliance, independent of AI. Section 3 outlines the principles and guidelines for ethical AI and is divided into five sub-sections: private initiatives for defining AI ethics principles (Sect. 3.1), IOs (with Sects. 3.2, 3.3, 3.4), and European Union (EU) (with Sect. 3.5). Section 4 provides an in-depth examination of the principles, and sub-principles concerning AI ethics. This analysis enables us to understand how the general principles of AI are applied in different sectors. Section 5 combines the general and sub-principles to create our reference model for ethical compliance with AI. Additionally, Sect. 5.1 provides practical applications of this model in the education sector. Section 6 suggests several recommended techniques to implement our AI ethical compliance model. Finally, Sect. 7 identifies some challenges in implementing a semi-automated software model to support the assessment of AI ethics compliance.

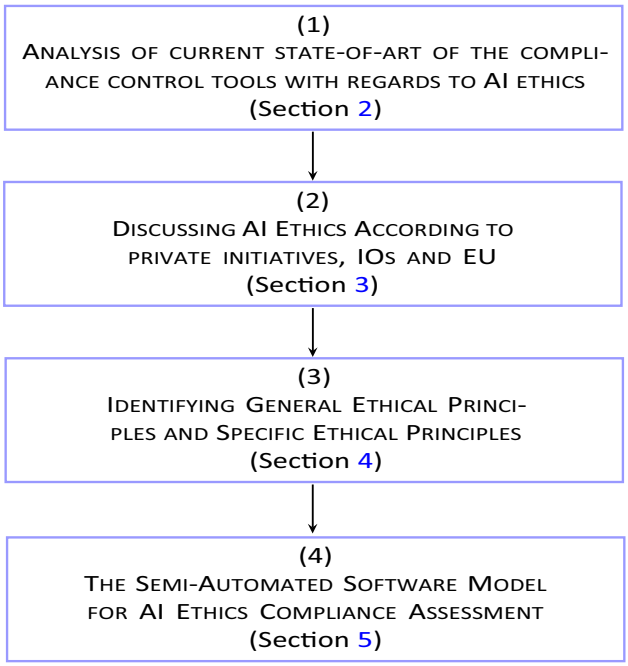

Fig. 1 Our approach and focus of the research

2 Related work

Several studies have developed many techniques for verifying a system's regulatory compliance (Sect. 2.1), and only a few of these studies focus on ethical compliance (Sect. 2.2).

Verification of a system's compliance can be performed manually or automatically. In this section, we review the techniques used to verify the ethical compliance of AI systems. It highlights the importance of ensuring that systems comply not only with regulatory standards, but also with ethical standards. This context provides the basis for understanding how ethical principles influence the design and implementation of algorithms and artificial systems, as discussed in subsequent sections.

2.1 AI ethics compliance review

The ethical compliance of a system can be checked through a manual process with some current techniques.

For instance, the Australian government has proposed a manual check using Australia's AI Ethics Framework [15]. The framework describes how the Australian AI Ethics Principles have been tested in a pilot project involving several of Australia's largest companies, including the Commonwealth Bank of Australia, Flamingo AI, Insurance Australia Group, Microsoft Conversational National, Australia Bank and Telstra. AI systems developed by companies have been evaluated through the ethical principles of AI. Microsoft has created the Microsoft Bot Framework, an open-source platform for designing chat bots. This system has been evaluated against the eight principles of AI ethics. The company started by responding to the question of whether the Australian AI ethical principles apply to conversational AI. The company conducted a thorough comparative analysis of Microsoft AI's conversational and ethical principles.

Vakkuri et al. [49] introduce the ECCOLA method for AI system developers and organisations. The ECCOLA approach enables assessment and integration of ethics into the AI development process. ECCOLA is based on a deck of cards that guides how to address ethical issues in AI. The method provides a toolkit, guidelines, and suggestions that developers can manually apply to ensure that the systems comply with AI ethics. ECCOLA is designed to be used throughout the development cycle by facilitating the educational aspect of AI learning ethics.

Brey et al. [9] presented the Sienna project, which provides a five-level model for translating ethical values into operational guidelines for building an ethical AI system. The approach facilitates the incorporation of ethical values at every stage of the development process by introducing them through concrete actions. Similarly, Umbrello et al. [46] have proposed an extension of the Value Sensitive Design (VSD) approach throughout the entire AI system design life-cycle, incorporating the principles of AI4SG (AI for Social Good) as VSD standards. The AI4SG-VSD method consists of four interactive phases, including context analysis, value identification, formulation of planning requirements, and prototyping. These two prototypical approaches to ethical AI design emphasise the importance of interdisciplinary collaboration between experts from various fields to integrate ethical values at each planning stage of the AI system development process. Gerdes [20] has highlighted that the difference between these two approaches is that the AI4SG-VSD method focuses on integrating AI-specific values and challenges, whereas the Sienna project provides a comprehensive framework for ethical AI design and deployment.

Research on the implementation of ethical standards in the design and development of AI systems is carried out mainly manually, and the number of manual approaches is also limited. In the Sect. 2.2, we aim to identify the automatic solutions presented in fields other than AI. This addition will provide valuable insights for designing our software model.

2.2 Compliance review

Hashmi et al. [21] explained the mining process to verify compliance with the regulatory requirements. This process involves scanning system logs to assess whether a company has met regulatory requirements during its business processes. Process mining allows to extract and collect log data about the activities, including information on the sequence where the activities are performed. The data is analysed to extract the process model, which helps to understand the actual process performance. Finally, this model is compared with regulation and reference standards to ensure accuracy.

Libal et al. [27] have developed the NAI Suite to verify the compliance of their product with GDPR. The NAI Suite uses automated deductions to simplify and automate GDPR compliance in AI by answering specific questions regarding compliance with Article 13 of the GDPR. The NAI Suite conducts an automatic compliance check by making specific deductions. The NAI suite facilitates the annotation of article 13 of GDPR, specifically with regard to the transparency requirements for data collection and processing enabling for the formalisation of the article. Once the article has been properly annotated, NAI Suite can respond to queries about GDPR compliance. Third-party users can use the NAI Suite to evaluate the GDPR compliance of their products or services. Questions can be generated automatically by a third-party tool and sent through API to NAI Suite for compliance

verification. Finally, the NAI Suite supports the ability to write questions directly in the tool. This function is mainly used for testing and as a support tool for lawyers and legal professionals.

Molinero et al. [31] have offered assistance to web developers in achieving compliance with web accessibility standards, presenting three automated tools: Watchfire Bobby, LiftNN for Dreamweaver, and Ramp. Watchfire Bobby, the first automated validation tool created, has essentially become *"de-facto standard"*. LiftNN for Dreamweaver receives the highest rating in a report that evaluated six tools for ease of use. At present, Ramp is the most recent validation tool available, and this particular product has little research to support it. Upon analysis, each tool generates reports showing the number of errors that occur on a website. In turn, these tools facilitate the identification and rectification of accessibility issues, rendering the website more accessible for users.

Nikkila et al. [33] described the embedded approach to rules, where the rules are embedded directly in computer instructions as part of the compliance checking system. The authors propose an infrastructure that allows automated checking of compliance with agricultural production standards. This infrastructure transfers the production standards in a computer-encoded and machine-interpretable format between stakeholders involved in contemporary agricultural production. The system is based on REST and is developed using the Ruby programming language and the Ramaze2 open-source Web application framework.

Some scholars have adopted a language-based approach, which represents normative knowledge as computable rules. Lee et al. [26] have developed Building Environment Rule and Analysis (BERA), a language for verifying digital building models and ensuring compliance with building regulations. Solihin et al. [41] also have explored this approach, creating the BIM Rule Language, a domain-specific language aimed at automating BIM26 rule checking. This language is concerned with defining a simplified schema in a relational database to represent a read-only building model that includes its geometry. Also, Preidel et al. [36] have presented the Visual Code Checking language. This language employs a formal language with visual syntax and visual semantics. It represents a modular system of signs and rules that use visual elements instead of textual ones on the semantic and syntactic levels. Finally, Milanovic et al. [29] used Business Process Integration that is based on the General Rule Markup Language. It translates rules and policies stated in different rule languages into one comprehensive rule language (REWERSE I1 Rule Markup Language—R2ML) and consistently processes them.

Amor and Dimyadi [2] have developed an automated compliance check system, named CORENET's ePlanCheck, that focuses on the relevant national codes for building control, barrier-free access, and fire safety. The system aims to provide an internet-based electronic submission platform for checking and approving building plans. The main component of the code-checking system is the FORNAX, an object library written in C++. Each object contains all relevant attributes for the Singapore codes and the corresponding regulations. Each object is designed to be extensible to meet the criteria of other countries. Similarly, Lee et al. [25] have designed the KBIM system to verify compliance with the Korean Building Act, the legislative authority for all construction activity in South Korea. KBIM makes use of KbimCode, the computable rules representation of the statutes' normative provisions. Furthermore, Clayton et al. [13] presented SMARTreviewTM APRTM which is an automated compliance check system on the building code. This system is integrated into the Building Information Modeling (BIM) software, and it is added to Autodesk's Revit BIM authoring tool that supports portions of the International Building Code. The author created an interface that helps the designer of the building to insert the information into the BIM model, initiate and inspect a review of plans, and produce reports. The system is innovative because it focuses on compliance with the fire safety standards of the International Building Code (IBC). Beach et al. [6] showed UpCodes AI BHR, which is a Revit add-on. It supports portions of the International Building Code, as well as several other codes applicable to various jurisdictions in the United States. These add-ons provide designers with compliance advice while working within the Revit software environment.

Dimyadi et al. [16] have developed ACABIM, a human-guided automation system that uses a workflow-driven approach guided by humans to automate tedious computable compliance tasks. This system enables human experts to concentrate on verifying performance-based designs that are more qualitative and require tacit knowledge, which machines cannot provide. It fully supports open standards and uses OpenBIM for sharing building information. In addition, it uses LegalRuleML to represent normative knowledge. LegalRuleML is an extension of RuleML designed to incorporate permissions and obligations for describing rules that reflect normative knowledge and account for significant legal and logical aspects.

Zhang et al. [51] employed the hybrid natural language processing (NPL) to extract normative knowledge from codes and standards. The aim is to assess compliance with the normative provisions during construction project design and implementation phases. In particular, the authors combine grammar and context-related aspects to automate the extraction of information from regulatory documents, and investigate the viability of two other techniques, i.e., phrase-structure grammar and dependency grammar, to extract information from complex sentences. Zhang et al. [52] used three techniques to automatically extract requirements from

regulatory documents in the construction industry and to enunciate them in a computer-processable format. These techniques comprise semantic modelling, semantic NPL techniques (including text classification and information extraction), and logical reasoning to facilitate automated analysis and processing of regulatory documents.

Some companies have created a regulatory compliance management tool called ZenComply [37]. The platform is intuitive, identifying areas of concern before compliance risks become significant threats. Then, it is scalable for small businesses and large enterprises as a Software-as-a-Service (SaaS) product. ZenComply's workflow management tools feature a single dashboard that documents the efficacy of the control in real-time, making compliance paperwork effective. Finally, the system aids in creating audit trails, enabling the back up of auditor questions' responses. Templates drive efficiencies in audit management, self-assessment, and vendor questionnaires.

We can summarise the current related work in Table 1, where the special features of the presented systems are described. The three characteristics that are mentioned include (i) the techniques used by the system creators, (ii) the type of compliance executed by the system (whether manual or automatic), (iii) the system's scope of application.

3 Review of principles and guidelines for ethical AI

In this section, we examine the fundamental ethical principles that guide the design and implementation of AI. These principles, derived from IOs' decisions and EU documents, provide a framework for ensuring that AI is used in an ethical manner that respects the values shared by society. This section builds a bridge between the examination of the ethical compliance of AI systems (Sect. 2) and the study of the principles that guide such compliance.

Our research focuses on examining the decisions adopted by IOs and EU regulations in order to identify key ethical principles essential for guiding the design and implementation of algorithms and artificial systems, ensuring their societal impact aligns with shared moral values. Through

Table 1 Description of the compliance control systems examined

System	Techniques	Test type	Scope
Australia's AI Ethics Framework	Ethical Compliance Checking Manual	Manual	AI Ethics
ECCOLA	Ethical Compliance Checking Manual	Manual	AI Ethics
Process mining	Mining data	Automatic	Regulation and reference standards
NAI Suite	Mining data	Automatic	Art. 13, GDPR
Watchfire Bobby LiftNN for Dreamweaver Ramp	Three automated evaluation tools	Automatic	Web accessibility standard
Infrastructure	REST implements Ruby programming language and Ramaze2 open source	Automatic	Agricultural production standard
BERA	Language based-approach	Automatic	Building regulations
BIM Rule Language	Language based-approach	Automatic	Building regulations
Visual Code Checking language	Language based-approach	Automatic	Building regulations
Business Process Integration	Language based-approach	Automatic	Building Regulations
CORENET's ePlanCheck	FORNAX (object library)	Automatic	Singapore's building regulations
KBIM system	KbimCode (computable rules representation)	Automatic	Korean Building Act
SMARTreview™ APR™	Integrated into the Building Information Modelling (BIM) software and is added to Autodesk's Revit BIM authoring tool that supports parts of the International Building Code	Automatic	Korean Building Act
ACABIM system	OpenBIM to share building information and LegalRuleML2 to represent normative knowledge	Automatic	Automating repetitive and computational tasks
A Hybrid natural language processing (NLP) system hybrid natural language processing (NLP) system to extract normative knowl- edge from provisions in codes and standards	Semantic modelling Semantic NPL techniques Logic reasoning	Automatic	Building regulation
ZenComply platform	Scalable Software as a Service (SaaS) platform	Automatic	Regulatory compliance in different sectors and business contexts

the exam of the international and European documents, we can define the concept of "machine ethics." This concept is not to be confused with other concepts such as "AI ethics," "robot ethics," or "computational ethics" from which it must be distinguished. As some scholars suggest, there are important differences between these notions. Segun [38] reviews the common terminology used in the field of AI ethics, highlighting the confusion and lack of clarity in the definitions of terms such as "machine ethics". In particular, the author argues that computational ethics represents a crucial frontier in the integration of ethics in autonomous intelligent systems because it involves experts in the programming of the systems, focuses on the design of algorithms that comply with ethical principles, addresses complexities such as ethical framing and moral uncertainty with experimental and procedural approaches, and, finally, involves interdisciplinary disciplines, fostering collaboration between ethics, logic, and computer science to develop ethically responsible AI. Stahl [42] also explores other conceptual distinctions, such as computer ethics and AI ethics, and proposes a new concept, the ethics of digital innovation ecosystems, toward which the author recommends that the debate evolve in the future. He suggests that understanding the similarities and differences between the two discourses can benefit them individually and lead to useful conclusions for larger socio-technical systems. He aims to shift the focus from specific technical artefacts, such as computers or artificial intelligence, to ethical issues that arise in the context of socio-technical systems. Finally, Mu¨ller [32] highlights the difficulty of establishing a coherent notion of machine ethics, arguing that to define what is meant by "ethics for machines" we would have to consider whether machines can have ethics in the proper sense of understanding how they would act according to ethical values or principles, or whether it is simply a matter of applying predetermined rules or algorithms without any real moral understanding.

This section is structured into two main parts. The first presents private initiatives to define ethical principles for AI (Sect. 3.1), and the second details international and European documents (Sects. 3.2, 3.3, 3.4, and 3.5).

The first part highlights the active role that scientists have taken in defining the ethical principles of AI, contributing to the creation of an inclusive ethical framework, and responding to the specific needs of the communities or sectors involved. These initiatives represent a common starting point from which to examine how these principles can be integrated or reflected in international and European policies. The second part underscores the importance of a formal and global approach to the definition of ethical principles for AI by international bodies and the EU. The bodies offer a broader and unifying vision that takes into account the global implications of AI and promotes the adoption of common principles that reflect universal values and ethical standards for AI. This dual perspective allows us to identify two complementary aspects in the process of defining ethical principles for AI.

3.1 Initiatives for defining AI ethics principles

There are many public and private initiatives that aim to define a list of ethical principles for AI. As Mittelstadt [30] points out, "to date, at least 84 such 'AI Ethics' initiatives have published reports describing high-level ethical principles, tenets, values, or other abstract requirements for AI development and deployment" (p. 1).

The approach of creating lists has been applied to medical ethics by Mittelstadt [30], who notes that many of the ethical principles converge on the four classic principles of medical ethics: respect for human autonomy, prevention of harm, fairness and clarity. The scholar notes that many of these initiatives produce vague principles and high-level value statements that promise to guide action, but in practice offer few specific recommendations and do not address fundamental normative and political tensions related to key concepts such as justice and privacy.

Several researchers have tried to define a set of shared AI ethics principles by examining AI guidelines. Jobin et al. [23] identified a global convergence around five ethical principles, including transparency, justice and fairness, non-maleficence, accountability, and privacy. However, while these principles are shared across different countries, their interpretation varies significantly. Fjeld et al. [19] noted that the study on AI principles documents revealed eight common themes, each consisting of three to ten principles. The foundational requirements for ethical AI, respecting human rights, include privacy, accountability, transparency, safety and security, fairness and non-discrimination, human control of technology, professional responsibility, and promotion of human values.

Anderson et al. [3], quoted by Mu¨ller [32], state that machine ethics is a field of study and discussion concerned with establishing ethical guidelines and principles for the behaviour of machines, particularly towards human users and, in some cases, toward other machines. Thus, machine ethics involves defining moral rules and norms that machines should follow in their functioning and interaction with humans and other automated systems.

Isaac Asimov's proposal of the "Three Laws of Robotics" [4] is an important starting point for the study of machine ethics and has left a permanent mark on thinking about how to ensure that automated systems behave ethically. The writer defined three laws defining that:

(1) A robot may not injure a human being or, through inaction, allow a human being to come to harm, (2) a robot must obey the orders given it by human beings except where such orders would conflict with the First Law, (3) a robot

must protect its own existence as long as such protection does not conflict with the First or Second Laws.

To these three laws is added a fourth, and (4) a robot may not harm humanity or, by inaction, allow humanity to come to harm. These laws set out the fundamental principles to be followed in designing an ethical machine: from the ethical principle of safety and prevention of harm to people in the practical applications of machine ethics, through responsibility and ethical interaction with human users, to the principle of ethical responsibility in the design and implementation of autonomous systems.

Asimov's ethical principles have been enriched over time thanks to the intervention of various international and supranational bodies such as the IOs and the EU.

These bodies respectively have a global and supranational perspective on AI ethics that transcends national boundaries. Their objective is to establish widespread and uniform definitions that encompass different nations and cultures. Therefore, as they represent different countries, this enables the states to implement common principles. Furthermore, IOs and the EU employ specialists who investigate and assess the particular impacts of AI on distinct domains. The use of experts confers credibility and legitimacy upon their guidelines. In addition, international and supranational decisions favour the process of harmonisation through the adoption of common principles or guidelines shared by member countries. In particular, the OECD, UNESCO and WHO are examples that help us understand how specific domains of expertise respond to the issue of ethics in AI. The three international organisations focus on three different areas, such as: economic, financial, scientific, social, environmental, training and development policy for the OECD; education, science, culture, communication, and information for UNESCO; and public health for the WHO. AI has made inroads into these fields, raising various ethical issues. This intrusion has led the IOs to intervene by defining guidelines and principles that can guide the ethical use of AI in that particular domain. Finally, the EU has been chosen because, following a tradition based on the civil law system, it has been regulating the ethics of AI since the beginning of the debate on AI since 2018, through a clear position on the development of ethical and legal frameworks [10].

Hence, in the next sub-sections, we examine some aspects of AI ethics guidelines issued by some IOs and the EU. Our analysis involves identifying commonly used words within these decisions and developing an ethical concept based on their guidelines.

3.2 The United Nations Educational, Scientific and Cultural Organisation

In the context of AI and ethics, on 25 June 2021, the United Nations Educational, Scientific and Cultural Organisation (UNESCO) drew up the *Draft text of the Recommendation on the Ethics of Artificial Intelligence* [48].

UNESCO states that the "AI system must be used appropriately and proportionately to achieve a particular legitimate purpose" (Rec. 26), and the selected AI approach should not violate fundamental values, especially by violating or abusing human rights.

AI method should be based on rigorous scientific evidence that is useful in assessing the validity, efficacy, and effectiveness of the AI system in the given context. This requires consulting published studies and research, laboratory experiments, and comparative analyses with other methods (Rec. 26).

UNESCO also encourages the control of safety and security risks that "should be addressed, prevented, and eliminated throughout the life-cycle of AI systems to ensure human, environmental, and ecosystem safety and security" (Rec. 27). Another important principle is to ensure that "the benefits of AI technologies are available and accessible to all, taking into consideration the specific needs of different age groups, cultural systems, different language groups, persons with disabilities, girls and women, and disadvantaged, marginalised and vulnerable people or people in vulnerable situations" (Rec. 28).

Thus, sustainable development of AI needs a "continuous assessment of the human, social, cultural, economic, and environmental impact of AI technologies" (Rec. 31).

UNESCO points out the importance of protecting users' privacy with respect to data collection and the social and ethical implications of such use. Additionally, it prioritises the 'privacy by design' strategy, expecting that data protection and privacy measures are incorporated into the system's design from the outset (Rec. 34).

Responsibility is a key principle, and member states should ensure the identification of individuals or legal entities accountable for all stages of the AI system's life-cycle. This allows us to determine those involved in the creation, management and use of the AI system and hold them responsible for the decisions made by the AI system (Rec. 35). UNESCO refers to both ethical and legal responsibility and liability of AI actors for decisions and actions based on the AI system in any capacity, based on their role in the life-cycle of the AI system (Rec. 42).

When algorithm-based decisions impact safety or human rights, people have the right to be fully informed and request explanatory details from the relevant AI actors or public sector institutions (Rec. 38).

Then, it needs to explain how an AI system makes decisions intelligible and provides insight into the outcome of AI systems. This includes "both the input, output and the functioning of each algorithmic building block and how it contributes to the outcome of the systems" (Rec. 40).

The operation and decisions of the AI systems should be recorded to be examined and analysed. Such measures ensure that AI systems can comply "with human rights norms and standards and mitigate threats to environmental and ecosystem well-being" (Rec. 43).

Finally, two additional principles are discussed, such as awareness and literacy that improve the understanding of AI systems. It is necessary to educate all individuals involved in the life-cycle of an AI system, including professionals and stakeholders.

3.3 The organisation for economic cooperation and development

In the area of AI ethics, the Organisation for Economic Cooperation and Development (OECD) issued a recommendation on 5 May 2019, outlining the *AI principles* [34].

The Recommendation identifies five complementary values-based principles for the responsible stewardship of trustworthy AI and calls on AI stakeholders to promote and implement them.

The first principle concerns human-centred values of fairness, whereby promotion of "values alignment" within AI systems. An AI system must adhere to ethical principles and human rights, allowing people to intervene and monitor to prevent inappropriate or dangerous actions (principle 1.2).

To uphold these values, the OECD promotes the use of safeguards such as Human Rights Impact Assessments (HRIA), Human Rights Due Diligence, Human Determination (i.e., Human in the Loop), codes of ethics, or quality labels and certifications aimed at promoting human-centred values and fairness. 'Human in the loop' refers to the presence of humans in decision-making involving AI systems, aimed at preventing the system from deciding on its own. All these tools help to assess in advance how the implementation of a system might affect human rights.

Two additional principles are transparency and explainability. Regarding transparency, the OECD outlines four activities: first, the disclosure of when AI is implemented, and the disclosure should be proportional to the importance of the interaction. Second, people should understand how an AI system is developed, trained, operated, and used in the relevant application domain. Third, provide meaningful information and transparent explanations regarding the nature of the information presented. Fourth, facilitating discussion among various stakeholders and potentially establishing specific bodies to promote general awareness and trust in AI systems (principle 1.3).

Explainability means enabling people affected by an AI system's outcome to understand how it was determined. This principle includes: (i) providing individuals affected by an AI system with easily understandable information that enables them to challenge the outcome; (ii) when AI actors explain an outcome, they may consider providing "in clear and simple terms, and as appropriate to the context—the main factors in a decision, the determinants, the data, logic or algorithm behind the specific outcome, or explaining why similar-looking circumstances led to a different outcome. This should be undertaken in a way that allows individuals to understand and challenge the outcome, while also respecting applicable personal data protection obligations […]" (principle 1.5).

AI system should not carry unreasonable safety risks, including those affecting physical security, under normal or foreseeable use or misuse throughout their lifespan. To ensure the reliable safety and security of the AI system, two approaches are suggested: (i) maintaining traceability and performing subsequent analysis and enquiry, and (ii) applying a risk management approach (principle 1.4).

Finally, considering accountability for organisations or individuals is essential. Accountability refers to ensuring the proper functioning of AI systems they design, develop, operate, or deploy, throughout their life-cycle and in accordance with their roles and applicable regulatory frameworks. They must also demonstrate this through their actions and decision-making processes.

3.4 The World Health Organisation

In the context of AI and ethics, the World Health Organisation (WHO) has underlined the importance of fairness, transparency, and accountability in the development and use of AI systems for the health sector.

The WHO released the *First Global Report on AI in Health and Six Guiding Principles for its Design and Use*, which was published on June 28, 2021 [50]. The purpose of these principles is to regulate and govern the use of AI in healthcare, minimising potential risks and maximising benefits. The protection of human autonomy is a fundamental principle, as humans need to maintain control of healthcare systems and decisions. This principle also underscores the need to protect privacy, ensure confidentiality, and obtain valid informed consent from patients within the appropriate legal frameworks for data protection.

One principle is to promote human well-being, safety, and the public interest by requiring AI technologies to comply with regulatory safety standards, accuracy, and efficiency in specified use cases. Quality control measures and continuous improvement in AI employment of AI enhance patient safety and well-being.

Transparency, explainability, and intelligibility are crucial principles set out by the WHO. They emphasise the publication or documentation of sufficient information before the design or deployment of AI technologies. This information should be easily accessible, facilitate meaningful public consultation, and encourage open debates on technology design

and its appropriate application. Responsibility and accountability are essential aspects of AI governance. Furthermore, stake- holders are responsible for ensuring the appropriate conditions and the participation of properly trained individuals when using AI technologies. Mechanisms should be available to address concerns and provide remedies for individuals or groups adversely affected by algorithm-based decisions.

Inclusiveness and equity are key considerations in the development of AI for healthcare. Technology must be designed to promote equitable access and use.

Finally, the WHO promotes the development of AI systems that are responsive and sustainable. This involves continuous and transparent assessment of AI applications to determine their adequacy and appropriateness in meeting expectations and requirements. Furthermore, encouraging the design of AI systems with minimal environmental impact, increased energy efficiency, and the ability to address potential workplace disruptions. One such example is training healthcare workers to adapt to the newly integrated AI systems.

3.5 European Union

We analyse EU three documents on ethical principles concerning AI. While IOs define general ethical principles and guidelines for AI, it is the responsibility of national and regional jurisdictions to develop specific regulations to implement these principles. The EU regulations explicitly guide the application of ethical principles to AI within the EU.

The EU has demonstrated its commitment to regulating AI through three fundamental regulatory acts, such as (i) Ethics guidelines for trustworthy AI, in 2019 (Sect. 3.5.1), (ii) the study panel examining the impact of the General Data Protection Regulation (GDPR) on AI in 2020 (Sect. 3.5.2), and (iii) the EU AI Act in 2020 (Sect. 3.5.3).

3.5.1 Ethics guidelines for trustworthy AI

The High-Level Expert Group on AI published the *Ethics Guidelines for Trustworthy AI*, on 8 April 2019 [22] defines seven key requirements and outlines ethical principles to be respected in AI use.

The first requirement involves human agency and supervision. AI systems should empower humans, allowing them to make informed decisions and promoting their fundamental rights by enlisting human-in-the-loop, human-on-the-loop, and human-in-command approaches.

The second requirement concerns the technical robustness and security of AI systems. These systems must be resilient and secure, as well as accurate, reliable, and reproducible. It is crucial to have a backup plan in place in case the AI system's security plan fails.

Privacy and data governance are essential to ensure complete respect for privacy and data protection. It is appropriate to have data governance mechanisms in place that take into account data quality, integrity, and legitimate access.

Transparency is also an important principle that must be maintained in data, systems, and AI business models through traceability mechanisms. AI systems and their decisions should be explained in a manner that is adapted to the relevant stakeholders.

AI systems can hurt people through discrimination and marginalisation of vulnerable groups. To prevent unfair bias, stakeholders should be involved throughout the entire life-cycle of these systems. AI systems should benefit all human beings, including future generations, while also taking into account the environment and other living beings. The social and ethical impact of AI systems should be carefully assessed.

The High-Level Expert Group on AI has finally requested accountability and reliability of AI results. To ensure that these principles are achieved, tools should be made available. Then, accountability and reliability require verifiability mechanisms to assess algorithms, data, and design processes. These mechanisms provide evidence in legal proceedings.

3.5.2 The impact of the general data protection regulation on AI

The EU has published *The impact of the General Data Protection Regulation (GDPR) on AI*, on June 25, 2020 [17].

The study notes the potential implementation of AI under the GDPR. However, the GDPR does not provide adequate guidance to data controllers; therefore, its provisions need extension and coordination to apply to AI. In any case, GDPR provisions relevant to AI include the processing of personal data (art. 4(1)), profiling (art. 4(2)), consent (art. 4(11) and art. 6), ensuring fairness, transparency (art. 5(1)(a)), purpose limitation (art. 5(1)(b)), data minimisation (art. 5(1)(c)), accuracy (art. 5(1)(d)), storage limitation (art. 5(1)(e)), fulfilling information duties (art. 13 and art. 14), providing information on automated decisions (art. 13(2)(f), 14(2)(g)), prohibiting automated decisions (art. 22(1)), allowing exceptions to the prohibition (art. 22(1) and 22(2)), handling automated decisions and sensitive data (art. 22(4)), and implementing data protection by design and by default (art. 25).

3.5.3 EU AI ACT proposal

EU has laid down the *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised*

rules on AI, on April 21, 2021 [18]. The European Parliament and Council reached political agreement on this proposal in December 2023, and Parliament finally adopted the EU AI Act on March 13, 2024.

This regulation aims to improve the operation of the internal market by establishing a consistent legal framework, primarily for the creation, advertising, and use of AI that conforms to Union values (recital 1). The proposal contains significant aspects that represent the main focus for future AI regulation.

The first guideline focuses on transparency in AI and includes several key elements. Regarding *transparency and information provision*, high-risk AI systems must be designed and developed in a manner that ensures that their operation is transparent enough for users to correctly interpret the system output. Providers must ensure that AI systems designed to interact with nature person are transparent regarding their nature as an AI system (art. 13). This is expected under the *transparency obligations for certain AI systems* (art. 52) and should be achieved unless the circumstances and context of usage make the AI system's nature obvious to the individuals involved.

The second specific guide concerns the definition of the term "Ethical," which denotes the adherence to the values of the EU and the Fundamental Rights outlined in the European Charter of Fundamental Rights. Therefore, the actions and behaviour of an AI system should not infringe on these values and rights. Some of the fundamental rights include respect for human dignity (art. 1), the right to life (art. 2), respect for private and family life (art. 7), protection of personal data (art. 8), freedom of expression and information (art. 11), freedom of assembly and association (art. 12), freedom of the arts and sciences (art. 13), and freedom to conduct a business (art. 16), protection of intellectual property (art. 17(2)), non-discrimination (art. 21), equality between women and men (art. 23), the rights of children (art. 24), the integration of individuals with disabilities (art. 26), the right to collective bargaining and action (art. 28), fair and just working conditions (art. 31), environmental protection (art. 37), the right to an effective remedy and a fair trial (art. 47), as well as the presumption of innocence and the right to defence oneself (art. 48).

Three categories of risk-based approaches have been identified: unacceptable risk, high-risk, and activity-specific risk.

An activity presents an unacceptable risk if it violates the values of the Union, including the infringement of fundamental rights. These activities, referred to as "prohibited AI practices," include: "(i) AI systems that deploy subliminal techniques beyond an individual's conscious awareness to significantly materially distort their behaviour, resulting in physical or psychological harm to that or another person. (ii) AI system that leverages the vulnerabilities of the group of individuals based on their age, and physical or mental disability to significantly distort the behaviour of someone within that group, causing physical or psychological harm to the individual or others. (iii) AI systems used by public authorities or on their behalf to assess or categorise the reliability of individuals based on their social conduct or personal characteristics" [18] (p. 12).

An activity is considered to be "high-risk" if the AI systems pose a significant threat to the health, safety, or fundamental rights of individuals. Such activities can be permitted on the European market only after fulfilling particular mandatory requirements and after a prior conformity assessment.

An AI system should be classified as high-risk when certain conditions are fulfilled, such as: (i) the "AI system is intended for use as a safety component of a product, or is itself a product, covered by the Union harmonisation legislation listed in Annex II" (art. 6(1)), and (ii) the product of which the AI system is a safety component, or the AI system itself as a product, is subject to third-party conformity assessment for "the placing on the market or putting into service of that product under the Union harmonisation legislation listed in Annex II" (art. 6(1)).

The proposal defines that specific AI systems must adhere to *transparency obligations* concerning their manipulation risks. These systems include those involving "(i) human interaction, (ii) detecting emotions or associating them with (social) categories using biometric data, and (iii) generating or manipulating content deep fakes" [18] (p. 16).

4 General and specific ethical principles for AI

Establishing principles and sub-principles responds to the issue raised by Mittelstadt [30] about the challenge of moving from principles to practice. The author discusses the lack of proven methods for translating high-level ethical principles into practical implementation, suggesting that such a process should adopt a multi-level approach. Central to this approach is the translation of abstract (and often contested) AI ethics principles into a set of "mid-level norms" and "low-level requirements." In the current context, our strategy is characterised as multi-level and incremental, focusing on defining the first two levels of principles while deferring the elaboration of low-level requirements. At this early stage of the project, the inclusion of low-level principles could potentially burden it, leading to unnecessary complexity and comprehension challenges.

Also, Taddeo et al. [44] develop a methodology for interpreting and applying the ethical principles of AI to specific practices, particularly in the context of the defence sector. The proposed methodology entails identifying the appropriate level of abstraction to model the AI life-cycle, taking into

account specific defence industry practices. The methodology then interprets the ethical principles of AI to derive specific requirements that must be followed at each stage of the AI life-cycle. Finally, the methodology defines criteria for purpose and context-specific harmonisation of ethical principles, ensuring that they are consistent with broad ethical goals and democratic values.

In this section, we identify and discuss: (1) general ethical principles and (2) the ethical sub-principles of AI, as gathered from our review of principles and guidelines for ethical AI.

4.1 General ethical principles

Thirteen general principles emerge from the analysis of the IOs and EU documents (Sect. 3) as shown in Table 2.

This is not a scale of universal ethical values for AI, but a list of principles and guidelines of ethical AI that are currently recognised in international and European decisions. Since universal principles are difficult to identify, our collection of thirteen data appears to be a combination of those identified through our analysis.

Mittelstadt [30] criticises the adoption of the medical ethics approach in the field of AI ethics. He believes that the list of ethical principles commonly used in medical ethics may not be directly applicable to the development of artificial intelligence (AI). This idea arises from the significant differences between the two sectors. The characteristics of AI, such as lack of common goals, fiduciary duties, established professional norms, and the absence of established methods to translate principles into practice, can make it difficult to achieve universal consensus on high-level ethical principles in AI.

Table 2 AI ethical principles extracted from the analysis of the IOs decisions

No	Principle
1	Safety and Security
2	Transparency and Explainability
3	Responsibility and Accountability
4	Privacy and Data protection
5	Fairness and non-discrimination
6	Human oversight and Determination
7	Inclusiveness and Equity
8	Sustainability
9	Responsive and Sustainable AI
10	Technical Robustness and Safety
11	Well-being and Safety
12	Awareness and Literacy human-centred values
13	Societal and Environmental well-being

4.2 Ethical sub-principles

The study of EU legislation leads us to realise that identifying general ethical principles for AI is only a first step in defining AI ethics, as these principles are broken down into different sub-principles depending on the domain in which an AI system is used. Depending on the context of use, principles, and sub-principles may come into play rather than others. For example, when implementing an AI system within the medical sector, its principles and their respective sub-principles would be different from those of an AI system implemented in the educational sector (Sect. 5.1).

Fjeld et al. [19] identified eight themes in their study of twenty AI documents, each containing three to ten principles. In total, they identified 47 sub-principles. In our study, we examined IOs and EU documents to check for the presence of the sub-principles identified by Fjeld et al. [19] in their research.

From our in-depth analysis of the international and European texts, we have identified sentences containing sub-principles by a method that combines the recognition of predefined keywords with the analysis of their context in the text.

The process starts with (a) the pre-processing of the text to make it ready for analysis, followed by (b) the identification of predefined keywords, and finally (c) the analysis of the context of the keywords in the text. First, we apply NLP techniques to optimise the text by removing non-essential elements and standardising the text representation for better analysis and identification of significant words. This includes removing irrelevant information such as punctuation and stop words, reducing words to their basic forms through stemming or lemmatization, and breaking text into smaller units such as sentences or tokens [8]. Text preprocessing techniques have been found to be a significant contributor to the accuracy of any text-based algorithm. Tabassum et al. [43] illustrate the importance of these preprocessing techniques in extracting meaningful information from text. The authors note that the use of NLP requires preprocessing techniques that are critical to the effectiveness of information retrieval systems because they affect accuracy and efficiency.

Once the text has been pre-processed and structured, the code searches a predefined list of keywords for their presence in the text. This task aims to simplify the effort of manually identifying terms from domain-specific text. In particular, keywords include principles and sub-principles, but also the variations of the latter, such as: “privacy by design”, “privacy design”, “by design”, “privacy by”. Two NLP techniques are used: lemmatization and synonym identification. While in the first case the expression is reduced to its basic form, in the second case the technique allows identifying synonyms associated with each principle.

Finally, we analyse the concordance, which "is a collection of the occurrences of a word-form, each in its own textual environment. In its simplest form it is an index. Each word-form is indexed, and a reference is given to the place of occurrence in a text" (p. 32) [40]. Specifically, we use the Keywords in Context (KWIC) technique that "is only one way of looking at corpus data, and that the definition of a concordance" [45]. In this context, we find all occurrences of each keyword in the text. For each occurrence we get the surrounding context, including the 5 words before and the 5 words after the keyword. The search window is set to 5 because it provides the opportunity to capture a context large enough to understand the circumstances in which the word is used. This technique allows us to understand how these words are used in the text and how they relate to surrounding words. In addition, concordance analysis can also reveal any synonyms or related words that may be used in the text instead of the standard keywords. This can be useful for a more accurate search for words of interest in the text. However, the KWIC technique has some limitations, such as not providing a complete understanding of the broader context in which these words are embedded. For example, a privacy keyword may appear in several sentences within a text, but its meaning may vary depending on the surrounding context. Without an understanding of the context, you risk drawing the wrong conclusions. To overcome the first limitation, we resorted to the supervision and control of legal experts.

Through our in-depth analysis of the regulatory texts, we have found, for example, that the UNESCO recommendation incorporate five sub-principles related to the privacy principle, including informed consent as an essential aspect of collecting personal data within the AI context (Rec. 33). Individuals must provide "explicit and informed consent" before their data is collected. Additionally, Rec. 33 establishes the right to erasure personal data and the control over its use in AI systems including collection, exchange, and use. Finally, Rec. 34 outlines the application of the privacy by design approach, stating that "algorithmic systems require thorough privacy impact assessments, including social and ethical considerations for their use".

Table 3 contains the themes and corresponding sub-principles identified by Fjeld et al. [19] in their study. A concise description of each sub-principle is provided, followed by the textual references of the IOs and EU documents where the concept of these sub-principles is mentioned.

5 The semi-automated software model for AI ethics compli ance assessment

Our approach involves several steps (Fig. 1). Based on the above reviews and analysis, this Section proposes the requirements for a semi-automated system for verifying ethical compliance of AI software.

Indeed, analysis of several IOs and EU documents presents an opportunity to identify the ethical principles and sub-principles regarding AI. These can then be combined to design a software model and its requirement for ethical compliance testing.

To summarise the stages in the construction of this model, we set up the following steps: (i) identification of general ethical principles for AI (Sect. 4.1). All thirteen principles are interrelated. For example, technical robustness and safety, well-being and security, transparency, and explainability are fundamental aspects of safety and security. Additionally, privacy and data protection guarantee fairness, non-discrimination, accountability, responsibility; (ii) identification of sub-principles (Sect. 4.2), and (iii) these principles and sub-principles serve for the design and requirements of the semi-automated software for ethical compliance of ethical software (Sect. 5 and Fig. 2).

Figure 2 illustrates a framework for assessing the ethical nature of an AI system. This assessment tool assists in evaluating the ethical nature of the system through two distinct levels of evaluation. The first level involves a general assessment to determine whether the AI system complies with general ethical principles. This initial assessment acts as a preliminary check. Any violation of the general principles deems the AI system unethical, eliminating the need for further evaluation. The second level entails a detailed evaluation of the AI system's compliance with the sub-principles specific to the sector in which it operates. Each respective sub-principle corresponds to those general principles.

For instance, a self-driving vehicle must adhere to principles such as safety and security. If meeting these standards, the vehicle has to comply with all three corresponding sub-principles, namely: safety, security, and security by design. In the healthcare sector, surgical instruments must comply with privacy regulations and potentially all corresponding sub-principles. Also, this tool has to adhere to accountability, but not necessarily all of its corresponding sub-principles. For example, it may not have to address environmental responsibility or meet evaluation and auditing requirements.

The proposed system can offer insights into why the AI system is labelled as unethical, highlighting the specific sub-principles that have been violated. As soon as the system detects a violation of an ethical principle or

Table 3 Overview of AI principles, sub-principles within the IOs and EU's documents

Principle	Sub-principle	Definition	Textual references
Privacy	Consent	Personal information of an individual cannot be collected, processed, or used by any third parties without their express consent	UNESCO Rec. on AI 33 WHO Report on AI: mentioned in protecting human autonomy value GDPR: art. 4(11), art. 6 EU AI Act Proposal: art. 5(1)(a), art. 5(3)
	Control over the Use of Data	People should be able to determine what information about them can be used, for what purposes, and in what contexts	UNESCO Rec. on AI 32 WHO Report on AI: mentioned in protecting human autonomy value EU AI Act Proposal: art.17(f)
	Ability to Restrict Processing	This value pertains to the entitlement to manage and restrict AI systems' exploitation of personal data	GDPR: art. 18
	Right to Rectification	People have the right to review their data and request corrections in the case of errors or omissions	GDPR: art. 16
	Right to Erasure	It entails the right to request the deletion of one's personal information from the entity that holds it	UNESCO Rec. on AI 33 GDPR: art. 17, art. 22(1)
	Privacy by Design	Developers and operators of AI systems should prioritise the privacy aspects, covering data collection, processing, storage, and sharing throughout the entire life-cycle of the system	UNESCO Rec. on AI 33 WHO Report on AI: mentioned in protecting human autonomy value GDPR: art.25 EUAIACTProposal: art.5(1)(b)
	Recommends Data Protection Laws	To ensure the protection of the individual's right to privacy, specific regulations and rules are necessary	UNESCO Rec. on AI 33
Accountability	Verifiability and Replicability	An AI system should be designed and implemented in a manner that protects against undesirable behaviour, preventing all forms of discrimination, bias, and other kinds of prejudice (Fjeld et al., 2020)	UNESCO Rec. on AI 43
	Impact Assessments	The potential effects of AI systems on society and human rights will be assessed by considering both the risks and threats to human rights, as well as the opportunities to safeguard and advance human rights [19]	UNESCO Rec. on AI 42 EU AI ACT Proposal: art. 9

Table 3 (continued)

Principle	Sub-principle	Definition	Textual references
	Environmental Responsibility	There is a responsibility to consider the ecological impact when implementing an AI system [19]	UNESCO Rec. on AI 18 WHO Report on AI: mentioned in promoting AI that is respon- sive and sustainable value
	Evaluation and Auditing Requirement	AI system has to be able to be audited. The outcome of that audit should be used to feed back into a system of checks and balances [19]	UNESCO Rec. on AI 42 EU AI ACT Proposal: art. 62
	Creation of a Monitoring Body	To create an organisation to make and monitor standards and best practices in the AI context [19]	UNESCO Rec. on AI 42 EU AI ACT Proposal: art. 53, art. 56
	Ability to Appeal	If an AI system makes a decision that affects a person, that person has the right to challenge that decision	UNESCO Rec. on AI 29 WHO Report on AI: mentioned in fostering responsibility and accountability value
	Remedy for Automated Decision	Biases arising from AI systems may be a consequence of the actions of designers, developers, builders, etc Mitigation mechanisms should be implemented to counteract any resulting harm [19]	UNESCO Rec. on AI 33 WHO Report on AI: mentioned in fostering responsibility and accountability value
	Liability and Legal Responsibility	Individuals or entities who are accountable for any harm caused by the AI system will be held liable	UNESCO Rec. on AI 42
	Recommends Adoption of New Regulations	AI systems must be designed and implemented in accordance with fundamental values and rights	UNESCO Rec. on AI 16
	Accountability Per Se	This value defines distinct levels of responsibility at each phase of the AI scheme [19]	UNESCO Rec. on AI 42 EU AI ACT Proposal: art. 17(1)(m)
Safety and security	Safety	AI systems require preventive safety measures during implementation to ensure safe and ethical operation Furthermore, even after implementation, the system needs to be monitored and corrective action taken to prevent harm	UNESCO Rec. on AI 27 WHO Report on AI: mentioned in promoting human well-being and safety and the public inter- est value? EU AI ACT Proposal: art. 5(1)(d)
	Security	It comprises testing the resilience of AI systems, sharing information on vulnerabilities and cyber-attacks, and safeguarding privacy, personal data integrity, and confidentiality. Data security can be achieved through anonymisation, de-identification, or aggregation	UNESCO Rec. on AI 27 OECD Rec. AI 1.4 WHO Report on AI: mentioned in ensuring transparency, ex- plainability, and intelligibility values
	Security by Design	Security should be considered from the outset of the AI system's design and development process. This facilitates the application of abstract concepts, such as general principles	UNESCO Rec. on AI 27

Table 3 (continued)

Principle	Sub-principle	Definition	Textual references
Transparency	Transparency	Developers of AI systems should prioritise transparency, clarifying the decision-making process, training data, and internal logic driving system behaviour	UNESCO Rec. on AI 37 OECD Rec. 1.3 WHO Report on AI: mentioned in ensuring transparency, ex- plainability, and intelligibility values GDPR: art. 5(1)(a), art. 13, art. 14 EU AI ACT Proposal: art. 11, art. 13, art. 14, art. 52
	Explainability	It refers to the ability of an AI system to explain how it reached a specific decision or conclusion clearly and understandably, enabling users to understand the rationale behind the AI system's decisions	UNESCO Rec. on AI 40 WHO Report on AI: mentioned ensuring transparency, explain- ability, and intelligibility values EU AI ACT Proposal: art. 13, art. 14
	Open Source Data and Algorithms	It consists of "developing common algorithms and promoting open research and collaboration to support technological advancement" [19]	EU AI ACT Proposal: art. 14
	Right to Information	Users have the right to receive information about the usage and potential interactions with the AI system clearly and objectively [19]	UNESCO Rec. on AI 38 EU AI ACT Proposal: art. 5(3)
	Notification when AI Makes a Decision about an Individual	The notification that an individual must receive upon the system reaching them is the subject of this discussion. This principle is related to transparency and accountability in the decision-making process	UNESCO Rec. on AI 38 EU AI ACT Proposal: art. 11, art. 16(h)
	Notification when Interacting with AI	People interacting with an AI system must be notified that their interaction is with an automated system	UNESCO Rec. on AI 38
	Regular Reporting	Companies should provide information about the use of AI systems to users. This information is necessary for transparency and accountability, allowing users to make informed decisions about their interactions with the system [19]	EU AI ACT Proposal: art.62
Fairness and non discrimination	Non-discrimination and the Prevention of Bias	It is necessary to reduce or eliminate bias in AI in both model training and technology implementation [19]	UNESCO Rec. on AI 30 OECD Rec. 1.2 WHO Report on AI: mentioned in ensuring inclusiveness and equity values GDPR: art. 22(1)(2) EUAIACTProposal: art.5(1)(b)

Table 3 (continued)

Principle	Sub-principle	Definition	Textual references
	Representative and High Quality Data	An AI system must be trained on data that represents the population it targets. Furthermore, the data must be of high quality to ensure effective learning by the system [19]	UNESCO Rec. on AI 16 OECD Rec. 2.2
	Fairness	AI systems must process data fairly and impartially to avoid generating discrimination in data management	UNESCO Rec. on AI 28 OECD Rec. 1.2 GDPR: art. 22(1)(2)
	Equality	This principle concerns the right of every person to use AI systems regardless of their socio-economic status, race, or religion, among others. All individuals should have access to the advantages and prospects that technology offers	UNESCO Rec. on AI 21 OECD Rec. 1.2
	Inclusiveness in Impact	The significance of giving equal access to and utilisation of AI systems to everyone, including those who have been excluded from benefiting from the technology, is underscored [19]	UNESCO Rec. on AI 21 OECD Rec. 1.2 WHO Report on AI: mentioned in ensuring inclusiveness and equity values
	Inclusiveness in Design	To develop AI systems that are inclusive, it is necessary to involve people from different backgrounds in the development of the AI system. This approach ensures, more diverse rights need to be taken into account in the development of AI systems	UNESCO Rec. on AI 21 OECD Rec. 2.4 WHO Report on AI: mentioned in ensuring inclusiveness and equity values GDPR: art. 25
Professional responsibility	Accuracy	It refers to the ability of an AI system to provide accurate answers or predictions derived from the data or models on which it has been trained	WHO Report on AI: mentioned in promoting human well-being and safety and the public inter- est values EU AI ACT Proposal: art. 15
	Responsible Design	Responsible design entails developing AI systems that are transparent in operation, while eliminating discrimination, protecting privacy, ensuring equity in outcomes, providing clear liability definitions in case of harm, and more	UNESCO Rec. on AI 42 WHO Report on AI: mentioned in fostering responsibility and accountability values EU AI ACT Proposal: art. 9
	Multi-stakeholder Collaboration	To implement an AI system, involving stakeholder groups is encouraged or required to ensure that these systems are developed and used in an ethical, fair, and socially responsible manner [19]	UNESCO Rec. on AI 47 EU AI ACT Proposal: art. 15(3)

Table 3 (continued)

Principle	Sub-principle	Definition	Textual references
Promotion of human values	Human Values and Human Flourishing	AI systems must be designed with consideration for human values and well-being. It is important to ensure the system respects these values at the design stage, as well as during and after implementation	UNESCO Rec. on AI 13 OECD Rec. 1.2
	Access to Technology	Granting access to AI to a wide range of individuals and ensuring that it brings benefits is a crucial ethical and human rights-friendly use of this technology [19]	UNESCO Rec. on AI 21
	Leveraged to Benefit Society	The use of AI and the exploitation of the benefits of AI should aim to serve the public interest	UNESCO Rec. on AI 22 OECD Rec. 1.1 EU AI ACT proposal: art. 11

sub-principle, the user receives notification of the infringement, with an indication of which ethical provision has been violated and why. This is done at two different moments: at system design and at runtime. In the latter scenario, an iterative method is established, allowing the validation script to periodically check the ethical compliance of the AI system. The checks will produce regular reports summarising the ethical compliance status of the AI system.

5.1 Applying our ethical compliance analysis model in education

This model can be applied to all domains, but we want to apply it to the education domain because it is a sensitive domain highlighting various ethical issues, including the protection of children and younger generations, who are considered vulnerable subjects whose rights must be protected in the face of the use of technologies. Klimova et al. [24] state that "despite the fact that they belong to the technologically savvy Gen Z and Millennials, [students] are still, or even more, vulnerable to the threats they are exposed to, such as surveillance or sexual harassment" (p. 6).

Chaudhry et al. [12] observe that after the Covid-19 there was a proliferation of different products and technological solutions for students. Despite that, many of these solutions lack the quality and robustness required to handle the considerable user demand, posing a major challenge to the ambitious field of educational technology. In addition, educational technology faces a critical issue in managing the large amount of data that serves as the foundation for artificial intelligence systems. The quality and effective management of this data is emerging as a critical issue. Some of the challenges have been identified by Akgun et al. [1] highlighting the various forms of prejudice and the ethical challenges of applying AI in educational settings in relation to students from kindergarten to 12th grade in the United States, covering an approximate age range of 5–18. Their focus is on the potential problems and social risks of AI applications with regard to privacy, surveillance, autonomy, bias, and discrimination. Privacy can be compromised by the exploitation of data through facial recognition systems. The student activities are continuously monitored through personalised learning systems and social networking sites. Autonomy can be compromised by jeopardising students and agency to manage their lives through predictive systems, and bias and discrimination by perpetuating gender and racial bias and social discrimination through automated scoring systems.

Finally, the application of AI in an educational context has been highlighted by UNESCO itself, which has provided guidelines and recommendations on the ethics of AI in education through the 2019 *Beijing Consensus on Artificial Intelligence and Education* [47] and the 2021 Guide for

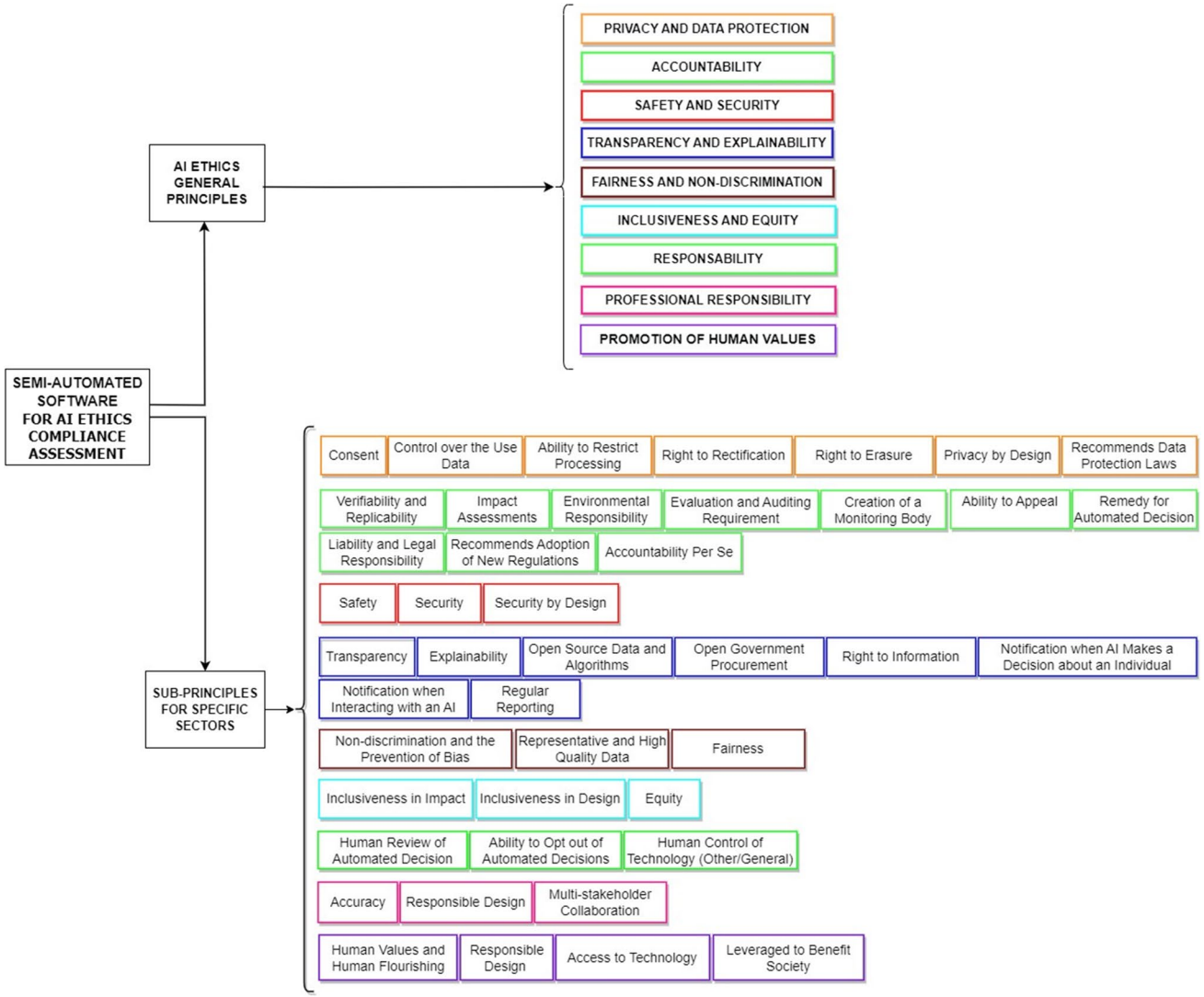

Fig. 2 Semi-automated software model and requirements for IA ethics compliance assessment AI Ethics

Policymakers on AI and Education [28]. These documents are essential references for defining the ethics of AI in education.

Our study applies the ethical compliance analysis model to an AI system used in the educational sector. Specifically, we focus on an AI-based collaborative learning platform for educators and students. Throughout our research, we apply our model to assess the ethical compliance of an AI system implemented in the educational process, focusing on two fundamental principles: privacy and accountability. Initially, we conduct a preliminary evaluation of *prima facie* compliance by examining whether the school has a privacy and accountability policy in place. If it does, we proceed to the second stage of evaluation, where we scrutinise whether the AI system conforms to the sub-principles relevant to privacy and accountability of that particular school. For instance, it checks the compliance with (a) consent, control over the use data, ability to restrict processing, right to rectification, right to erasure, privacy by design, recommend data laws protection, and (b) verificability and replicability, impact assessment, evaluation and auditing, requirement, liability, and legal responsibility. The assessment process flow diagram is presented in Fig. 3.

Our model performs a second-level assessment as described in Table 4. The monitoring system for ethical compliance poses specific questions for each sub-principle. Table shows an analysis of two key principles, such as privacy and accountability.

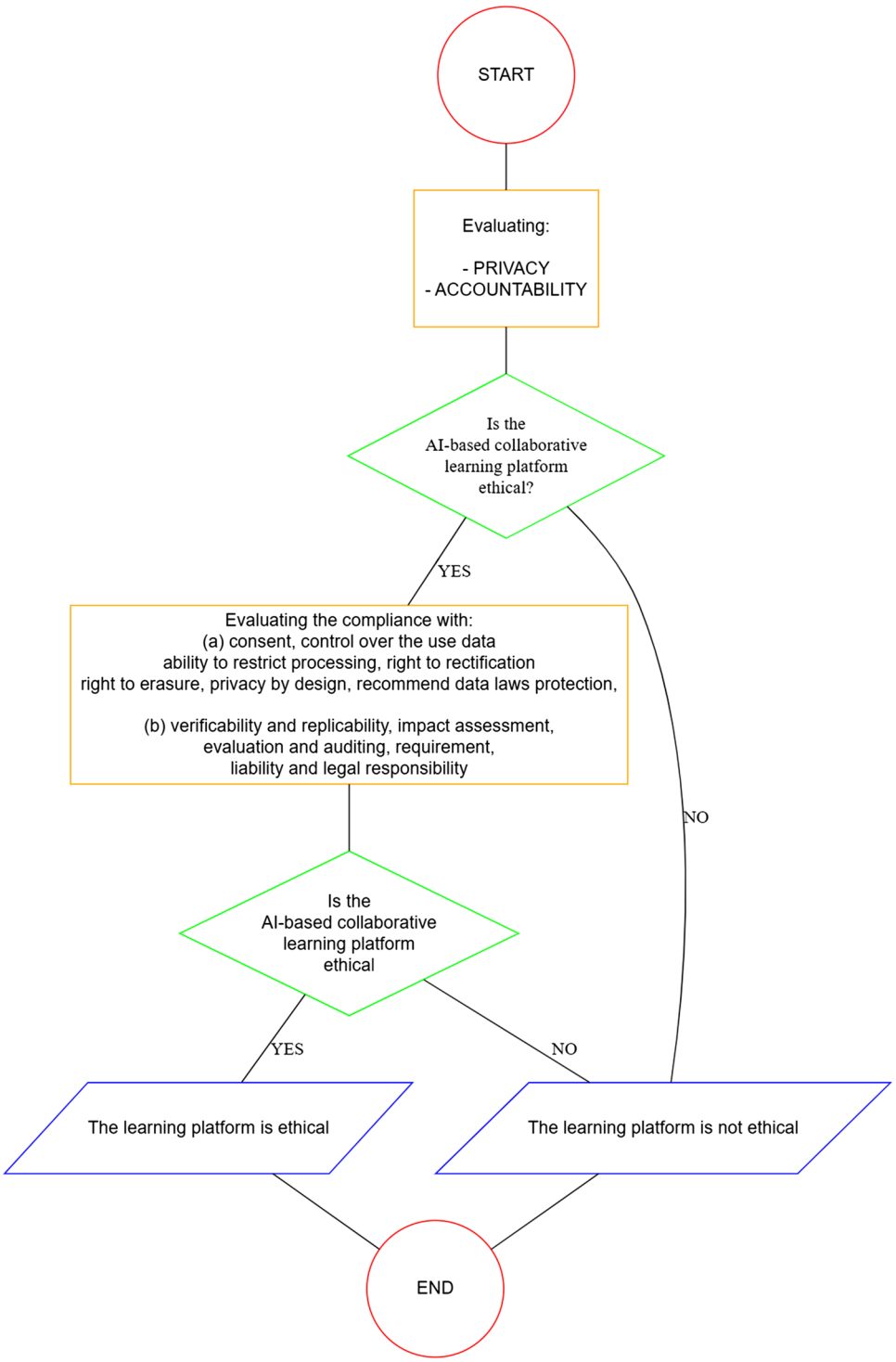

Fig. 3 Workflow of the semi-automated software model

6 Techniques for building our software model for AI ethical compliance assessment

To implement the proposed model, we propose certain techniques for its construction. The model could be implemented through an interactive platform that integrates transparency, privacy, and other algorithms to evaluate the compliance of the AI system with general principles. Users communicate via the platform through an interface that allows them to analyse the ethics of the system from different angles. The platform provides them with results in the form of charts and graphs so that they can understand them. Users can also provide feedback to improve the system's capabilities.

Table 4 Applying our semi-automated software model in education sector

Principles	Sub-principles	Application
Privacy	Consent	*Are parents and students asked for consent before personal information is collected or used?*
	Control over the use of data	*Are parents or students aware of how personal information will be collected and used by the AI- based collaborative learning platform? Do par- ents or students have the possibility to change their mind at any time, e.g., by opting out of the use of AI, or by choosing what information to share on the platform?*
	Ability to restrict processing	*Can parents and students prevent certain in- formation from being shared on the platform?*
	Right of Rectification	*Can relatives or students rectify personal data that are collected by the AI platform?*
	Right to Erasure	*Can relatives or students request the deletion of personal data, including the account, chats,* etc.*?*
	Privacy by Design	*Are security standards and data encryption respected? Are systems regularly updated to protect against vulnerabilities,* etc.*?*
	Recommended Data Protection Law	*Does the platform comply with privacy laws?*
Accountability	Verifiability and Replicability	*Are parents, students, and teachers able to understand how the data is used by the AI system? Can they see the documentation of the algorithm to be able to analyse it?*
	Impact Assessment	*Is the platform capable of modifying its algorithms as a result of the evaluation it has of user learning and decisions? Is the system able to collect user feedback?*
	Evaluation and Auditing Requirement	*Is the platform regularly audited?*
	Liability and Legal Responsibil ity	*Do the contract and terms of use specify who is responsible for any damages arising from the use of the platform?*

A second proposal is the creation of a chatbot that checks the ethical compliance of the AI system. The chatbot will be trained to recognise ethical situations using machine learning techniques. The chat will then be integrated with business intelligence tools to provide detailed reporting and data analysis. The chatbot will be able to communicate not only with the user providing the input, but also with the AI system. The user can then provide feedback to the chatbot. Such a system can be used both at design time, in an ethics-by-design perspective, guiding and helping the AI software designer in her choices, and afterwards for checking and assessing the AI ethics compliance of the resulting software.

The third proposal is to create a system that combines ontology with semantic rules. In this case, knowledge about the ethics of AI is organised in a coherent and detailed way, and rules are created that allow us to arrive at the ethics of a logical system by inference.

However, experts are expected to examine the triad of proposals and assess the reliability of the results produced by the systems. Subsequently, the results of this evaluation should be reviewed by a group of AI ethics experts.

7 Challenges in implementing a semi-automated software model to support AI ethics compliance assessment

A semi-automated software model to support the ethical assessment of AI must overcome several challenges before it can be implemented. We highlight just a few of these to provide some perspective. The **first challenge** regards the question of who should have the power and authority to validate the alignments of AI systems, there are various perspectives. An approach is to consider whether governments should be involved. Another option is for individual companies to establish internal mechanisms for ethical verification. Alternatively, an independent third party could be responsible for this task. In the case of an independent third party, questions arise about how it should be formed and the extent of power it should possess.

There is indeed no single solution, because it may be necessary to combine the efforts of govern ments, businesses and citizens, recognising that different sectors may require different regulatory approaches. The principle of subsidiarity, which encourages decision making at the most local and effective level, can be a helpful reference point. In the context of AI, this could mean that decisions on ethical review and adaptation are taken at the level where they can be most efficiently and appropriately addressed.

However, each intervention has its own advantages and disadvantages. Governments can bring a standardised

approach to AI ethics and alignment, but the approach is more bureaucratic, and the process of creating and implementing regulation can take longer. Companies could establish their internal ethical review and alignment processes, promoting self-regulation within the sector. However, they may prefer regulation to the detriment of ethics in order to achieve personal goals, resulting in a lack of transparency. Establishing an independent third-party body, such as an independent ethics evaluation commission, may be desirable, but the impartiality elements should be defined, starting with the appointment of the members of the commission. The qualifications and skills required for the ethical evaluation need to be defined, and it may be necessary to have an ad hoc specialist for each area.

The **second challenge** is the reluctance of companies to be transparent. To assess the ethics of an AI system, it is often necessary to understand how it works, its algorithms and how it makes decisions. Commercial companies are often reluctant to provide detailed ("white box") or application programming interface (API) access to their AI systems, unless required by law. Such corporate reticence limits the ability to monitor and ensure ethical behaviour in AI systems developed and deployed by private companies. As a result, it is difficult for regulators or enforcement agencies to accurately verify that software is being developed and used ethically. Taking the appropriate action is not straightforward. One solution could be to outline rules that can ensure transparency by forcing companies to be transparent by disclosing the details of their algorithms. However, transparency is a principle that must be balanced with other principles, such as intellectual property and the protection of inventions.

The **third challenge** is the risk of not reducing bias and unfairness in AI systems. This would require the adoption of an objective approach that can identify and mitigate them, as well as appropriate metrics and evaluation techniques that could be used to ensure the ethical and impartial operation of the AI system. However, the adoption of these techniques is not an easy task, as Pagano et al. noted [35]. In a systematic review, the authors show that it is difficult to determine which fairness metric is most appropriate for a given use case. This is because for a given use case, different fairness metrics do not produce consistent results, leading to different assessments of what is fair or equitable in a given context. This leads us to believe that there is no universal solution or single model architecture that is always the best to ensure fairness in all cases. Rather, you need to carefully evaluate the specific characteristics of your use case and adopt a model architecture and fairness metrics that best fit these characteristics.

The **fourth challenge** is to ensure that the software is also updated with respect to the changing nature of what is considered ethical over space and time. The concept of ethics is subjective because the meaning attributed to what is ethical cannot be determined in a universal and immutable way, but can vary considerably between cultures, historical, and social environments. In the field of ethics for AI a number of new concepts are emerging such as "machine ethics", "robot ethics", or "computational ethics", "ethics of digital ecosystems", etc. (see the discussion in Sect. 3). To face this challenge, it may be necessary to implement mechanisms for automatically updating the software to stay up-to-date with emerging developments and discussions on AI ethics.

The **fifth challenge** is to consider how the system can be concretely adaptable to different geographical areas characterised by specific cultural norms and ethical values, so that the system can perform effective evaluations of the ethics of an AI system that is adapted to different contexts. Beyond these challenges, our contribution advances existing knowledge on ethical compliance systems for the following reasons. Most of the current proposals for AI ethics compliance, reviewed in Sect. 2, aims at providing compliance checking of software. This means once code is developed, we assess whether it is compliant to some AI principle or regulation. Most companies put a special effort into data privacy verification, but as far as the AI ethics principles are concerned, all the principles listed in Fig. 2, should be taken into account. The specificity of our proposal lies in the following points:

- Compliance assessment is modular and flexible and covers all AI ethical principles.
- It allows a special focus on some of the principles depending on the application domain (see example in Table 4).
- Our tool can be used both to assess the software (after it has been developed) as well as before in an ethics-by-design approach. In that perspective, it can serve to further shape and design the AI software as a result of the assessment.
- The proposal of using a chatbot to interact whether with a human or the software itself renders the whole assessment process more fluid and user-friendly.

8 Conclusion

This paper arises from the need to define the current state of the art in the ethical evaluation of AI systems. Our hypothesis suggests that AI ethics evaluation systems are scarce, especially automatic, or at least semi-automatic systems, inherent to the AI sector. AI software like any other software needs to be dependable as defined by Avižienis et al. [5]: "the dependability of a computing system is the ability to deliver service that can justifiably be trusted". In their seminal paper, dependability's attributes of software systems include availability, reliability,

safety, confidentiality, integrity, and maintainability. Not surprisingly, most of these attributes are covered by the ethics principles we identified, in particular safety, confidentiality, integrity, and reliability. In the specific case of AI software, this early notion of dependability needs to be completed with additional principles such as transparency and explainability, or fairness and non-discrimination.

The review of recent works confirms that the majority of existing systems are inherent to a domain other than AI, and for AI they mostly have manual evaluation tools, such as internal company regulations. The study of this aspect is particularly important for companies producing AI systems, as they face uncertainty regarding a system's ethical compliance. Uncertainty has led companies to define their internal regulations, ensuring the development of systems that comply with the ethical principles identified by companies based on international and regional regulations. However, this process remains rather laborious due to the diversity of principles and guidelines for ethics AI and the difficulty in understanding which regulation to apply, especially given the transnationality of products placed on the market.

To address this issue, we present a software model and its working process which serve as the basis for the implementation of software that can both assess the AI ethical compliance of AI systems after their development but as well as in a prior development and ethics-by-design perspective.

The proposed model operates on two levels of evaluation. The first level is a preliminary to the second, in the sense that if the first step is not passed, the evaluation process stops. The first level aims to carry out a superficial evaluation in terms of compliance with the general ethical principles of AI, assessing, for example, the existence of policies adopted by companies to apply the ethical principles. The second evaluation has a specific focus on ensuring that AI models comply with the relevant sub-principles. It is important to note that the control system assesses only those sub-principles that are relevant to the specific sector in which AI is implemented.

Our future work will involve a detailed implementation of the system and the incorporation of the necessary techniques. Going forward, there is a crucial need for software to promote the responsible and trustworthy use of AI in various domains.

Author contributions Centre Universitaire d'Informatique (CUI), Université de Genève. Maria Assunta Cappelli and Giovanna Di Marzo Serugendo.

Funding Open access funding provided by University of Geneva. This research was supported by Innosuisse (Schweizerische Agentur für Innovationsförderung) Innovation Cheque within the framework of the innovation project n. 62509.1 INNO-SBM, entitled: "Reconciling companies and AI regulations initiative".

Declarations

Conflict of interest No conflicts.

References

1. Akgun, S., Greenhow, C.: Artificial intelligence in education: addressing ethical challenges in K-12 settings. AI Ethics **2**(3), 431–440 (2021). (**ISSN 2730-5953 2730-5961**)
2. Amor, R., Dimyadi, J.: The promise of automated compliance checking. Dev. Built Environ **5**, 100039 (2021)
3. Anderson, M., Anderson, S.L.: Machine ethics: creating an ethical intelligent agent. In: Machine Ethics and Robot Ethics, pp. 237–248. Routledge (2020) . (**ISBN 9781003074991**)
4. Asimov, I.: Runaround: a short story. In: Astounding Science Fiction, March 1942, p. 1940. Gnome Press, New York (1950)
5. Avižienis, A., Magnus, V., Laprie, J.-C., Randell, B.: Fundamental concepts of computer system dependability. In: IARP/IEEE-RAS Workshop on Robot Dependability: Technological Challenge of Dependable Robots in Human Environments, Seoul, Korea, 21–22 May 2001, pp. 1–13 (2001)
6. Beach, T.H., Hippolyte, J.-L., Rezgui, Y.: Towards the adoption of automated regulatory compliance checking in the built environment. Autom. Constr. **118**, 103285 (2020)
7. Bessen, J., Impink, S.M., Seamans, R.: The cost of ethical AI development for AI startups. In: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 92–106 (2022)
8. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python (2009). https://tjzhifei.github.io/resources/NLTK.pdf
9. Brey, P., Dainow, B.: Ethics by Design and Ethics of use in AI and Robotics. The Sienna Project-Stakeholder-Informed Ethics for New Technologies with High Socioeconomic and Human Rights Impact (2021). https://sienna-project.eu/digitalAssets/915/c_915554-l_1-k_sienna-ethics-by-design-and-ethics-of-use.pdf
10. Cappelli, M.A.: AI CODE, LIVRE 1: SOFTware/HARDlaw. In: Georg, Comtesse, X. (eds.) Partie 1: Première série d'entretiens (pp. 12-17) - Partie 2: Deuxième série d'entretiens (pp. 56-63) (2021)
11. Cappelli, M.A., Di Marzo Serugendo, G.: RecOnCiling Companies and AI Regulations InitiAtives (ROCCIA) (2023). https://archive-ouverte.unige.ch/unige:172701
12. Chaudhry, M., Kazim, E.: Artificial intelligence in education (Aied) a high-level academic and industry note 2021. SSRN Electron. J. (2021). https://doi.org/10.2139/ssrn.3833583
13. Clayton, M., Fudge, P., Thompson, J.: Automated plan review for building code compliance using bim. In: Proceedings of 20th International Workshop: Intelligent Computing in Engineering (EG-ICE 2013), pp. 1–10 (2013)

14. de Laat, P.B.: Companies committed to responsible AI: from principles towards implementation and regulation? Philos Technol **34**, 1135–1193 (2021)
15. Department of Industry, Innovation, and Science. Australia's artificial intelligence ethics framework. https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles (2019). Accessed 3 Sept 2023
16. Dimyadi, J., Clifton, C., Spearpoint, M., Amor, R.: Computerizing regulatory knowledge for building engineering design. J. Comput. Civ. Eng. **30**(5), C4016001 (2016)
17. European Parliamentary Research Service Scientific Foresight Unit (STOA). The impact of the general data protection regulation (gdpr) on ai. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRSSTU(2020)641530 EN.pdf (2020). Accessed 25 June 2023
18. Europena Commission. Proposal for a regulation of the European parliament and of the council laying down harmonised rules on AI (AI act) and amending certain union legislative acts, com (2021) 206 final - 2021/0106 (cod). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN (2021). Accessed 25 Oct 2023
19. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M.: Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication (2020)
20. Gerdes, A.: A participatory data-centric approach to AI ethics by design. Appl. Artif. Intell. **36**(1), 2009222 (2022)
21. Hashmi, M., Governatori, G., Lam, H.-P., Wynn, M.T.: Are we done with business process compliance: state of the art and challenges ahead. Knowl. Inf. Syst. **57**(1), 79–133 (2018)
22. High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (2019). Accessed 8 Oct 2023
23. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019)
24. Klimova, B., Pikhart, M., Kacetl, J.: Ethical issues of the use of AI-driven mobile apps for education. Front Public Health **10**, 1118116 (2023). https://doi.org/10.3389/fpubh.2022.1118116
25. Lee, H., Lee, J.-K., Park, S., Kim, I.: Translating building legislation into a computer- executable format for evaluating building permit requirements. Autom. Constr. **71**, 49–61 (2016)
26. Lee, J. K.: Building environment rule and analysis (BERA) language and its application for evaluating building circulation and spatial program. PhD thesis, Georgia Institute of Technology (2011)
27. Libal, T.: Towards automated GDPR compliance checking. In: Trustworthy AI-Integrating Learning, Optimization and Reasoning: First International Workshop, TAILOR 2020, Virtual Event, September 4–5, 2020, Revised Selected Papers 1, pp. 3–19. Springer (2021)
28. Miao, F., Holmes, W., Huang, R., Zhang, H.: AI and Education: Guidance for Policy-Makers. Guides, Manuals and Handbooks. UNESCO (2021). https://doi.org/10.54675/PCSP7350
29. Milanovic, M., Kaviani, N., Gasevic, D., Giurca, A., Wagner, G., Devedzic, V., Hatala, M.: Business process integration by using general rule markup language. In: 11th IEEE International Enterprise Distributed Object Computing Conference (EDOC 2007), pp. 353–353. IEEE (2007)
30. Mittelstadt, B. D.: AI ethics - too principled to fail? CoRR, abs/1906.06668 (2019)
31. Molinero, A.M., Kohun, F.G., Morris, R.: Reliability in automated evaluation tools for web accessibility standards compliance. Issues Inf. Syst. **7**(2), 218–222 (2006)
32. Müller, V.: Ethics of Artificial Intelligence and Robotics (2020). https://plato.stanford.edu/entries/ethics-ai/?utm_source=summari
33. Nikkila, R., Wiebensohn, J., Nash, E., Seilonen, I., Koskinen, K.: A service infrastructure for the representation, discovery, distribution and evaluation of agricultural production standards for automated compliance control. Comput. Electron. Agric. **80**, 80–88 (2012)
34. Organisation for Economic Co-operation and Development. Recommendation of the council on artificial intelligence. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449 (2019). Accessed 10 Oct 2023
35. Pagano, T.P., Loureiro, R.B., Lisboa, F.V.N., Peixoto, R.M., Guimaraes, G.A.S., Cruz, G.O.R., Araujo, M.M., Santos, L.L., Cruz, M.A.S., Oliveira, E.L.S., Winkler, I., Nascimento, E.G.S.: Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big Data Cogn. Comput. **7**(1), 15 (2023)
36. Preidel, C., Borrmann, A.: Automated code compliance checking based on a visual language and building information modeling. In: ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction, volume 32, page 1. IAARC Publications (2015)
37. RiskOptics. What is compliance automation? https://reciprocity.com/resources/what-is-compliance-automation/ (2022). Accessed 15 Sept 2023
38. Segun, S.: From machine ethics to computational ethics. AI & Soc. **36**, 263–276 (2021)
39. Shahriari, K., Shahriari, M.: IEEE standard review — ethically aligned design: a vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In: 2017 IEEE Canada International Humanitarian Technology Conference (IHTC), pp. 197–201 (2017)
40. Sinclair, J.M.: Corpus, Concordance and Collocation. Oxford University Press, Oxford (1991)
41. Solihin, W., Eastman, C.: Classification of rules for automated BIM rule checking development. Autom. Constr. **53**, 69–82 (2015)
42. Stahl, B.: From computer ethics and the ethics of AI towards an ethics of digital ecosystems. AI Ethics **2**, 65–77 (2022)
43. Tabassum, A., Patil, R.R.: A survey on text pre-processing & feature extraction techniques in natural language processing. Int. Res. J. Eng. Technol. (IRJET) **7**(06), 4864–4867 (2020)
44. Taddeo, M., Blanchard, A., Thomas, C.: From AI ethics principles to practices: a teleological methodology to apply AI ethics principles in the defence domain. SSRN Electron. J. (2023). https://doi.org/10.2139/ssrn.4520945
45. Tribble, C.: What are concordances and how are they used? In: O'Keefe, A., McCarthy, M. (eds.) The Routledge Handbook of Corpus Linguistics, pp. 167–183. Routledge, New York (2010)
46. Umbrello, S., Van de Poel, I.: Mapping value sensitive design onto ai for social good principles. AI Ethics **1**(3), 283–296 (2021)
47. UNESCO. Beijing consensus on artificial intelligence and education. In: International Conference on Artificial Intelligence and Education, UNESCO (2019). https://unesdoc.unesco.org/ark:/48223/pf0000368303. Accessed 29 Mar 2024
48. United Nations Educational, Scientific and Cultural Organisation. Recommendation on the ethics of artificial intelligence (2021). https://en.unesco.org/about-us/legal-affairs/recommendation-ethics-artificial-intelligence. Accessed 4 Sept 2023
49. Vakkuri, V., Kemell, K.-K., Jantunen, M., Halme, E., Abrahamsson, P.: Eccola: a method for implementing ethically aligned AI systems. J. Syst. Softw. **182**, 111067 (2021)
50. World Health Organization. Who issues first global report on artificial intelligence (AI) in health and six guiding principles for its design and use (2021). Accessed 20 Oct 2023
51. Zhang, J., El-Gohary, N.: Extraction of construction regulatory requirements from textual documents using natural language

processing techniques. In: Computing in Civil Engineering, pp. 453–460. ASCE (2012)
52. Zhang, J., El-Gohary, N.: Information transformation and automated reasoning for automated compliance checking in construction. In: Computing in Civil Engineering, pp. 701–708. ASCE (2013)