DOI: 10.47743/eufire-2023-1-4

OPERATIONALISING ARTIFICIAL INTELIGENCE ETHICAL PRINCIPLES IN BUSINESS – A CONCEPTUAL FRAMEWORK

ALEXANDRU CONSTANTIN CIOBANU
Alexandru Ioan Cuza University of Iași
Iași, Romania
alex.ciobanu.msft@outlook.com

GABRIELA MEȘNIȚĂ
Alexandru Ioan Cuza University of Iași
Iași, Romania
gabriela.meșniță@feaa.uaic.ro

Abstract
Historically, technology was always a key differentiator and an enabler for business or societal developments. With the rapid advancements in technology, businesses have been able to increase their efficiency, productivity, or profitability through automation and digitization. The European Union elaborated clear strategies that aims overall economic development trough digital transformation, one of the core technology components being Artificial Intelligence (AI). While the added value of AI technologies, in terms of optimization, efficiency, automation is clear and undeniable, there are different challenges related to the ethical aspect of AI use. Hence the AI ethics discussions are currently very present in the public space. A collection of different AI ethical frameworks was redacted either by the researchers or by governments or by the tech industry, but developing consistent AI ethical system is still a grey area. Some of the main challenges revealed by different researchers are related to the practical implementation of ethical principles in AI technologies during its lifecycle.
This paper proposes a conceptual framework for operationalizing AI ethical principles in business contexts. The framework is based on a comprehensive literature review of existing AI ethical frameworks, which highlights the current gaps in their implementation. The proposed framework addresses these gaps by providing a step-by-step approach that can be easily integrated into existing business processes. It covers various stages of AI development, including problem formulation, data collection and preprocessing, model training, model evaluation, and deployment. The practical validation of the framework will be conducted in future work. The results suggest that the proposed framework can provide a systematic approach to operationalizing AI ethical principles in business contexts, thereby contributing to the development of responsible AI systems.
Keywords: *artificial intelligence; ethics, framework; practice; business.*
JEL Classification: M15.

1. INTRODUCTION

The study of ethics applicable to Artificial Intelligence technologies is still in an early stage that presents unexplored opportunities from a scientific point of view. At the same time, it is observed that as companies adopt AI more and more, defining and promoting ethical principles related to AI is widely seen as one of the best ways to ensure that AI does not cause unintended harm.

Since the early stages of AI development, researchers have expressed concern about the ethical implications of its use in society (see Turing, 1950; Wiener, 1954). As AI technologies have developed, it has been found that they can have a tangible and significant impact on people's daily lives, such as influencing decision-making, improving quality of life, automating routines, and so on. As a result, the discourse on AI ethics has moved out of the academic enclave and entered the consciousness of the public and decision-makers (Barn 2019).

Therefore, there has been a rapid proliferation of ethical documents and declarations based primarily on principles, frameworks, standards, and codes of conduct proposed by the AI development industry, academia, or governmental and non-governmental organizations (national, European or global). European Comission has defined clear priorities and rules related to the development and use of AI. Therefore, they redacted a *Proposal for AI regulation* that sugest "*adopting a human-centric approach for digital technologies including artificial intelligence"* (European Comission, 2021). It has become increasingly clear that, although the existence of these documents may be necessary for creating the pro-ethical conditions so necessary (Floridi, 2019), the practical implementation of ethics at the AI level is far from being achieved (Vidgen *et al.,* 2020).

This is because many ethical principles dedicated to AI frameworks cannot be clearly implemented in practice, as demonstrated by some research (Haas, Gießler and Thiel, 2020; Morley, Elhalal and Garcia, 2021). There are studies that show that although ethical frameworks can be effective for promotion campaigns, they fail to manage the ethical operationalization of AI, with the risk of causing some of the damages they are intended to prevent (Ville *et al.,* 2019; Burt, 2020). Previous research (Floridi *et al.*, 2018; Morley *et al.*, 2023) shows that from the perspective of operationalizing an AI ethical framework, the businesses need to find solutions to a series of issues related to ethical management of an AI or to the specific limitations and motivation that AI developers may have when trying to implement AI ethical principles. Nevertheless, there is still a grey area when it come about measuring and tracking the ethical or unethical decisions of an AI.

To ensure that AI brings the anticipated added value, we must first comprehensively and interculturally understand the potential ethical risks related to AI development. Hence an AI ethical framework should combine both elements of ethical theory (such socio-cultural aspects) and technical elements.

Our research aims to develop an ethical framework applicable to AI that targets the operationalization of ethical principles from the moment of development to the implementation and post-implementation of an AI technology. Additionally, it is important to note that, in the scientific approach, we consider ethics to be more of a *continuous process*, rather than just a task.

The initial research questions that lead us to the development of our AI ethical framework were: *How ethical principles or moral norms can be created at the level of the AI algorithms or AI training models? How can we ensure that the ethical norms we want to define are essentially correctly understood by a software developer or an AI practician? How can we ensure that AI technologies can handle possible deviations from some ethical norms?* This paper does not claim to have a priori answer to these questions, but through the chosen methodology and the proposed AI ethical framework, we aimed to obtain concrete results that will allow interested companies to have ethical control and integrated management when implementing AI technologies.

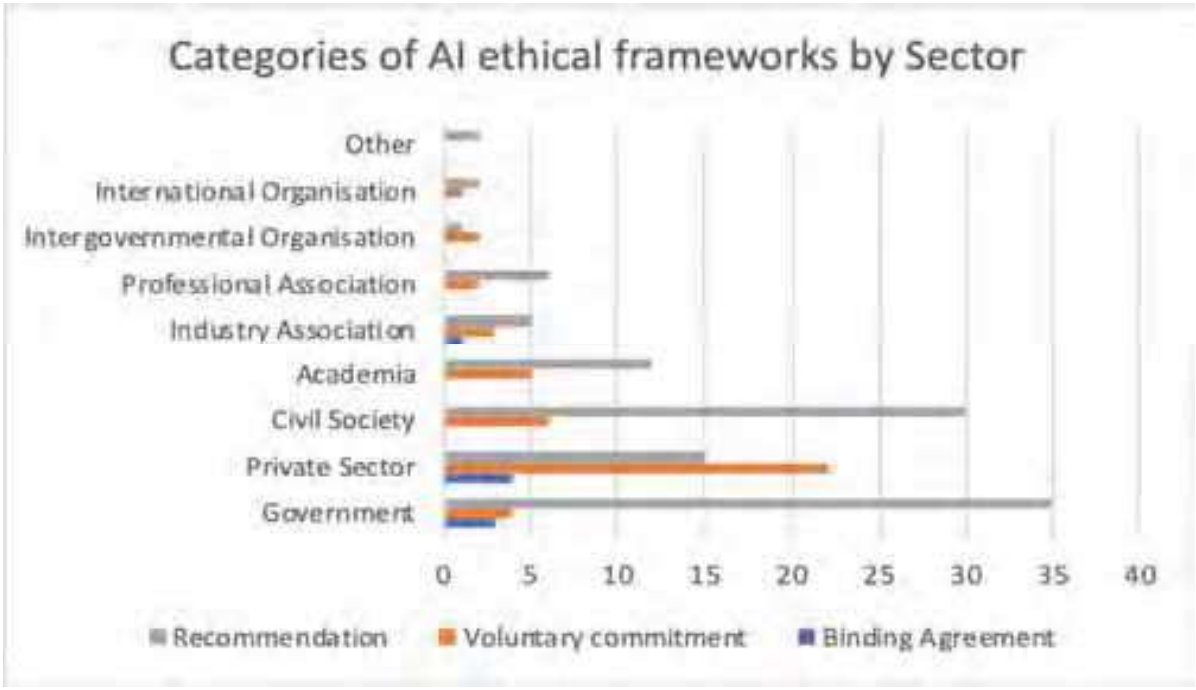
2. LITERATURE REVIEW

Through an examination of the current literature, it has been observed that numerous recent publications have highlighted a notable deficiency in many ethical frameworks applicable to artificial intelligence (AI), namely, the absence of essential considerations regarding how ethical values can be effectively implemented in practice (Hagendorff, 2020; *Morley et al.,* 2020). Furthermore, in addition to the lack of operationalization of ethical norms within AI ethics frameworks, some researchers have also noted the absence of guidance regarding the potential consequences - including legal, compliance, security, and other ramifications - of possible ethical deviations resulting from AI decisions (Jobin, Ienca and Vayena, 2019; Mittelstadt, 2019; Hagendorff, 2020).

In conducting a review of literature on the subject, we have identified a multitude of documents created by diverse stakeholders, including technology industry producers (such as Google, IBM, Microsoft, Amazon, etc.), governments (such as those behind the Montreal Declaration, the UK House of Lords Select Committee on AI, the European Commission's HLEG European Council, UNESCO, OECD, and others), and academic institutions (such as the Future of Life Institute, IEEE, AI4People, Japanese Society of AI, Oxford Digital Labs, among others). The authors of these documents seek to formalize presented principles (Anderson and Anderson, 2021) by proposing normative constraints (Turilli, 2007) on what an AI technology can and cannot do in society.

However, the existent frameworks do not always provide clear directions for the practical operationalization of ethical principles at the technical level.

Furthermore, according to Algorithmwatch.org, in 2019 (see Figure 1), it is shown that ethical guidelines for AI did not have enforcement mechanisms.



Source: Algorithm Watch (2023)

Figure 1. **Distribution of the AI ethical frameworks by its issuers**

Out of 160 documents presented in the AlgorithmWatch.org database, only 10 have practical enforcement mechanisms. Both private sector and public sector policies are mostly voluntary commitments or general recommendations. Surprisingly, the private sector relies largely on voluntary commitments, while government actors make recommendations for administrative institutions. Many of the AI ethics guidelines contain a wording that minimizes the scope of the document, presenting them as proposals.

A consensus has emerged among academic, industry, and governmental stakeholders that the operationalization of AI ethics should serve as a reference point for communicating expectations and evaluating the results and effects of AI. Nevertheless, challenges remain in this regard. As Hagendorff (2020) notes, although nearly all existing AI ethics guidelines propose technical solutions, few provide technical explanations. Consequently, AI practitioners may encounter difficulties in operationalizing the sometimes-abstract ethical principles at the algorithmic level (Calvo and Peters, 2019). The gap between ethical principles and their practical implementation is considerable and influenced by factors such as complexity, variability, subjectivity, and a lack of standardization, including variable interpretation of the "components" of each ethical principle (Alshammari and Simpson, 2017).

Our research underscores the importance of a complete (end-to-end) approach when discussing the significance of an ethical framework applicable to AI technologies. Research findings (Morley, Elhalal and Garcia, 2021)

demonstrate that AI practitioners possess an abstract and relatively narrow understanding of ethical principles and how they can be translated into practice. This suggests that AI practitioners are primarily motivated to translate ethical principles into practice to comply with legislative requirements. As Morley, Elhalal and Garcia (2021) note, this appears to be the sole justification for investing additional resources in AI product design. Furthermore, legislation in this area is not yet consistent and fails to keep pace with changes in social norms or attitudes, which can occur quite rapidly (especially in the digital environment).

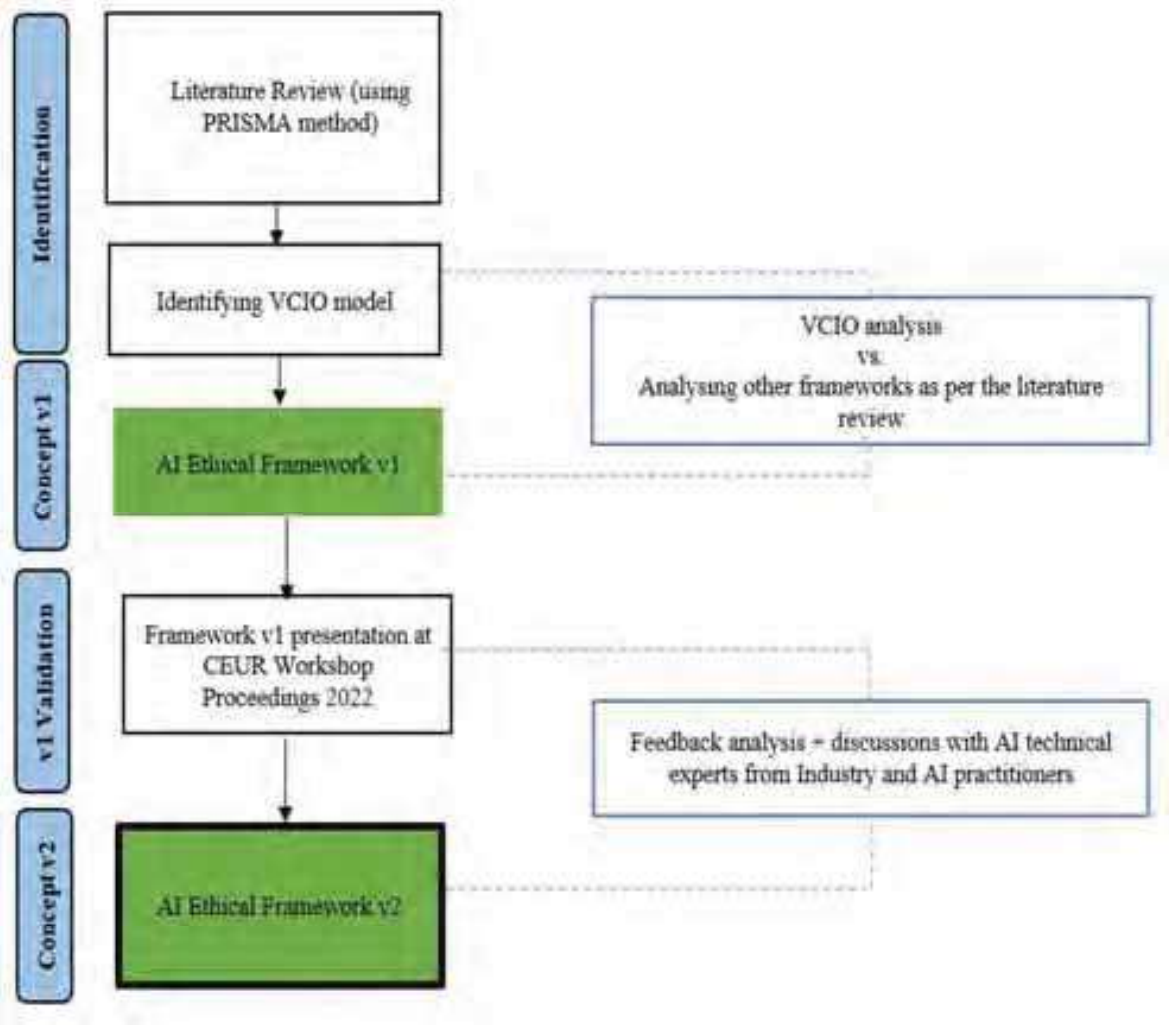
Building upon the aforementioned aspects, our research proposes an ethical framework applicable to AI technologies that has the capability to operationalize and transpose ethical principles from theoretical into the practical implementation at the level of AI practitioners. Recent studies (Ayling and Chapman, 2022) emphasize the importance of developing/analysing an ethical framework from three perspectives: *evaluating the impact of the framework, the ability to audit the framework, and the availability of technical tools within the framework for designing AI ethics.*

There are diverse approaches to what an ethical framework should or should not contain, just as there are different prescriptive approaches to what an ethical AI technology should mean. For this paper, we considered designing an AI ethical framework that should bring *added value* within a business organization that want to implement AI ethical technology. The approach of the ethical design of an AI technology centered on added value (for individuals, society, etc.) is not new (Stephanidis *et al*., 2019), but it seems to be a necessary condition for operationalizing the ethical principles applicable to AI.

3. METHODOLOGY

The research method that was adopted to create the ethical framework we propose is illustrated in Figure 2. As a starting point we considered the analysis of the existing scientific literature as described above. For that we used PRISMA method and that helped us to analyses ethical principles and ethical guidelines applicable to AI. Based on this analysis we have identified the VCIO ethical guide (Hallensleben *et al*, 2020) that has a practical description of the implementation of ethical principles in AI technologies based on *Values, Criteria, Identifiers and Observables*. One important note is that based on the analysis of other related works on ethics, we concluded that, although the VCIO ethics guide makes a step forward from the perspective of operationalizing AI ethics, it lacks on technical elements related on how ethical principles must be managed at the datasets or AI training models level. Thus, we proposed the addition of our vision to complete the VCIO ethical framework, aiming to create an ethical framework that addresses both the normative aspects of ethical principles and the technical elements specific to practical implementation at the level of AI technologies. This first version of the ethical framework that we proposed obtained a first scientific validation in the

CEUR Proceedings I-ESA 2022 conference (Ciobanu and Meșniță, 2022). Starting from the obtained feedback, we conducted a series of discussions with AI experts and AI practitioners after which we obtained the final version that we propose in this research paper.

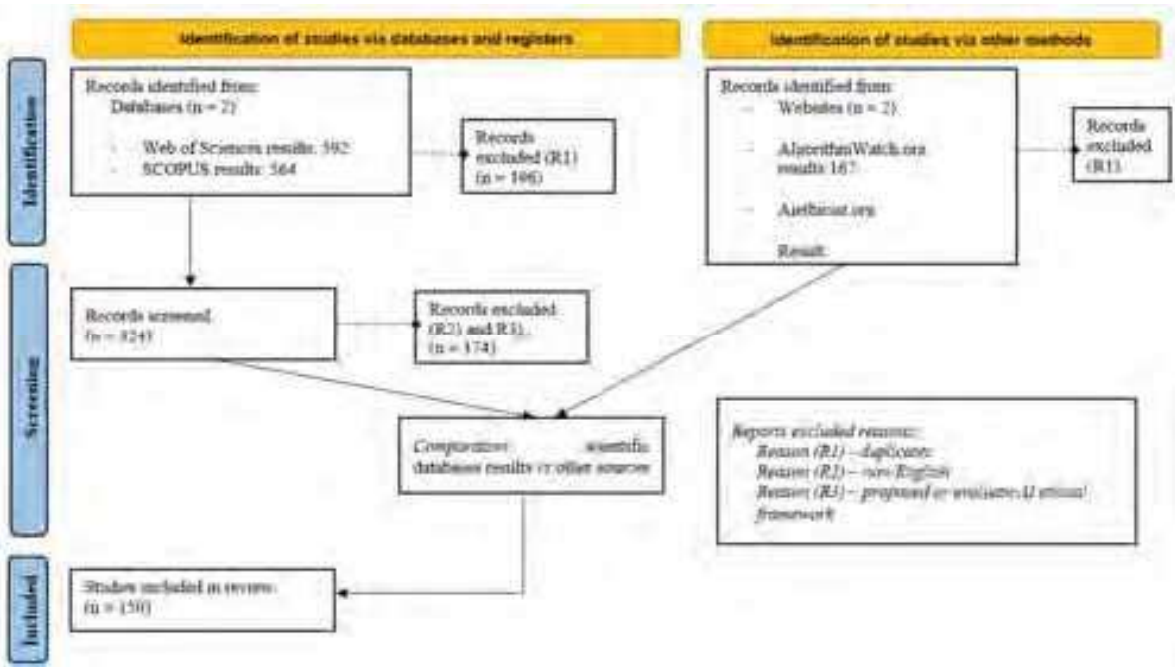


Source: self-representation

Figure 2. Methodology design to identify the proposed AI ethical framework

3.1 Using PRISMA method to identify the relevant AI ethical frameworks

The PRIMSA method was used as per Figure 3, to have a more detailed picture of other research that addresses a) the ethical issues that can be imposed on AI technologies and b) the current state of the practical applicability of existing ethical frameworks that facilitate the development, implementation and monitoring of AI technologies.



Source: self-representation

Figure 3. The use of PRISMA method to identify the most relevant AI ethical frameworks

This way we have identified that the main challenge of AI ethical frameworks are one related to the operationalisation of the ethical principles in practice. At the same time, this first methodological approach also highlighted the requirements that must be met in order to move from the normative approach to the need for ethics in AI to how we achieve the practical implementation of ethical principles in an AI technology. Also we identified the VCIO model (Hallensleben and Hustdet, 2020) which stands out as a good guide candidate that can (in a first phase) translate ethical principles from the perspective of AI applicability.

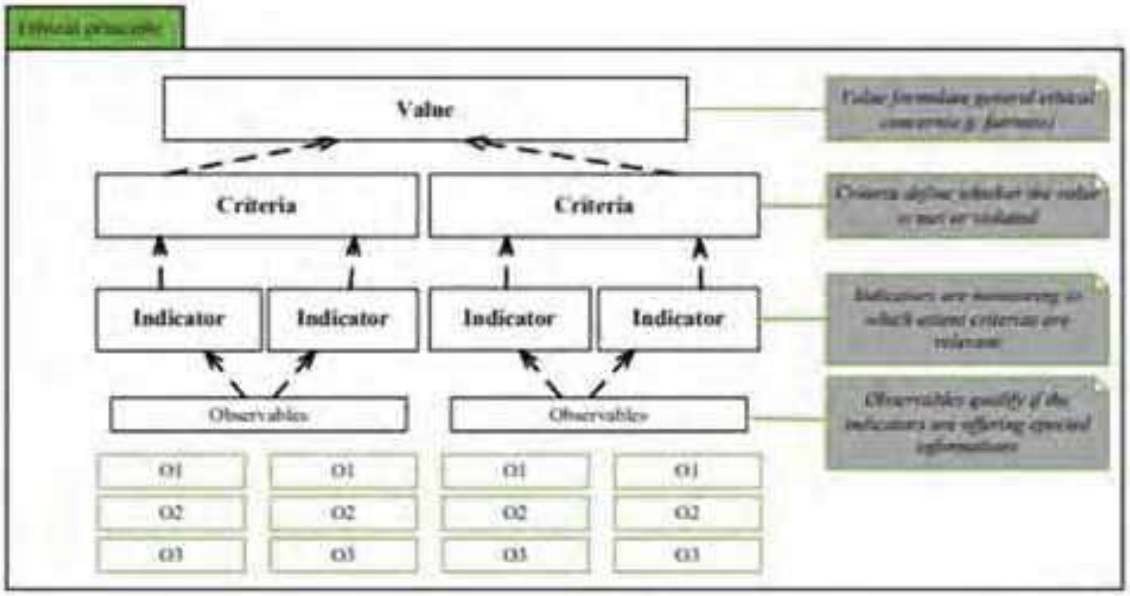
3.2 The VCIO AI Ethical model

The VCIO model is the result of a joint effort of an interdisciplinary group of experts called AI Ethics Impact Group (Hallensleben and Hustdet, 2020). By comparison with other ethics guidelines studied for this paper, VCIO brings as a novelty the need to measure ethical principles according to the field of applicability of an AI technology. Prescriptive ethical values, such as transparency or non-discrimination, are understood in different ways by different people or industries. This leads to uncertainty within organizations developing AI systems, while also hampering the work of AI regulators. The lack of specific and verifiable principles thus undermines the effectiveness of ethical guidelines.

The VCIO model highlights four essential components for operationalizing and evaluating ethical principles applicable to AI: *values, criteria, indicators and*

observables. Ethical principles are identified as abstracted values (similar to the subjectivism specific to the definition of ethical principles). The VCIO model framework shows that it is essential to have other components to perform these tasks. Thus, the criteria, indicators and observables can be used in defining the values that are proposed.

In order to practically implement AI ethics, the VCIO model comprises 4 levels (see Figure 4) that allow the operationalization or translation of ethical principles from theoretical norms into practical implementation tasks.



Source: adapted from AI Ethics Group Impact (2020)

Figure 4. High level functions of the VCIO model

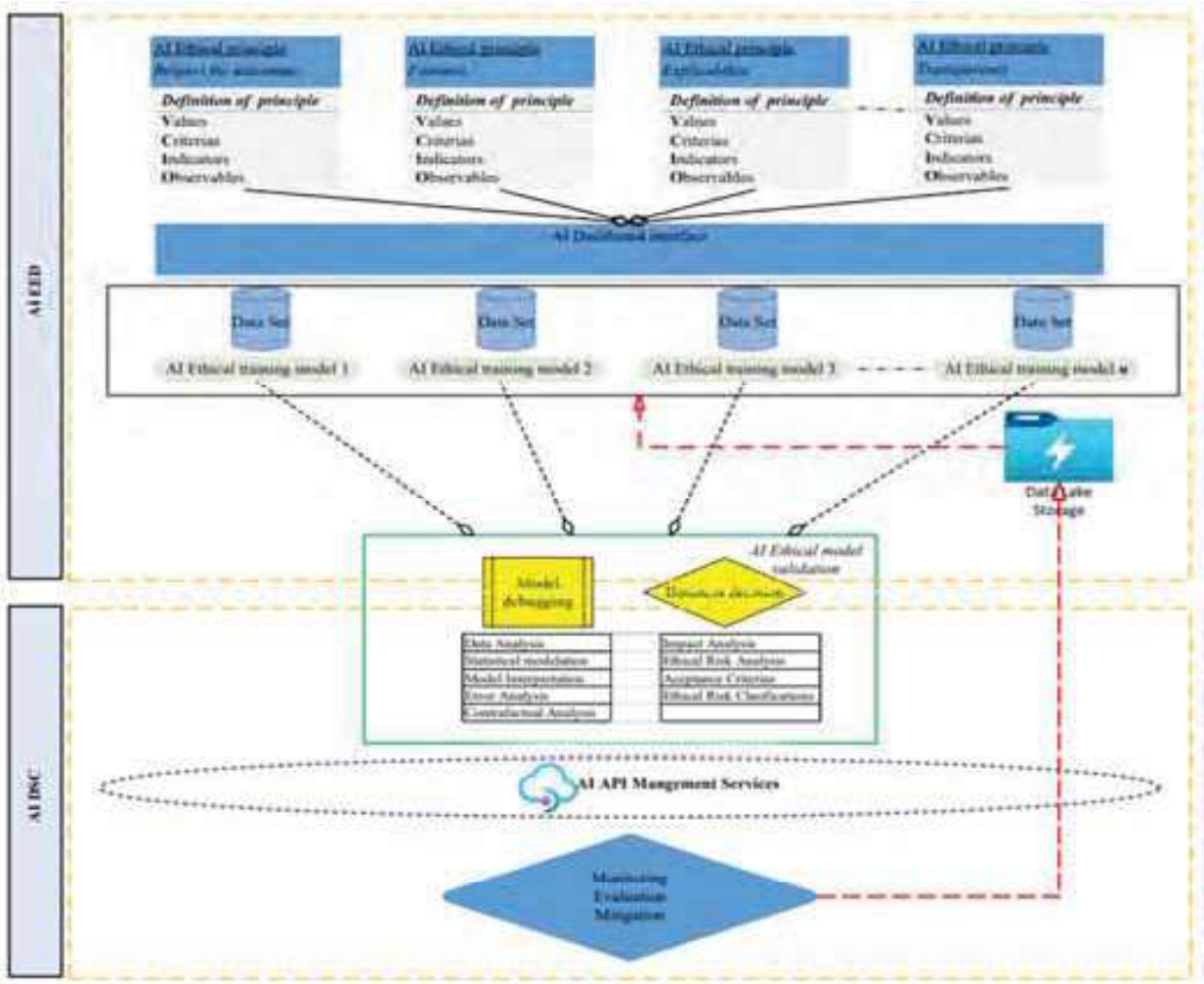
Summarizing, we considered the VCIO model as a starting point for the framework we propose, because it can define the requirements necessary to achieve a certain value that is identified with ethical principles. From a scientific point of view, the model can help operationalize ethical principles by concretizing general values and by breaking them down into measurable criteria, indicators and observables. On the other hand, there seems to be a lack of argumentation of the technical details that should be considered for the translation from *a)* ethical principles defined by values and concretely measured by criteria, indicators and observables *to b)* their implementation at the algorithmic level in the training models that are the basis of AI. In the framework we proposed VCIO framework could be seen rather as tool or instrument that can be used in the initial phase of explaining what an ethical principle should mean for an AI technology.

Hence, we are suggesting a unified AI ethical framework that includes both the methods by which we can translate the ethical values (according to the VCIO

model) into the training models, and a detection and updating mechanism once ethical deviations are detected in the initial algorithms.

4. OUR APPROACH FOR AN AI ETHICAL FRAMEWORK FOR BUSINESS

In this section of the research, we will describe the components of the proposed ethical framework (v2) applicable to AI technologies implemented in business organizations. As a result of the methodological approach detailed in the previous section, this framework v2 is an updated version of the framework v1 (Ciobanu and Meșniță, 2022) and brings additions related to the technical control elements of the training models, as well as functions related to the decision-making model related to the ethical aspects relevant to the business, as a result of the implementation of AI. (See Figure 5).



Source: self-representation

Figure 5. An AI Ethical framework for business organizations

The framework is divided into two conceptual sections connected through API (Application Programming Interface) systems, through which we suggest a holistic approach to operationalizing an AI system that addresses challenges from both AI producers and consumers.

4.1 AI Embedded Ethics By Design (AI EED)

AI Ethics By Design is not a new concept in the field of AI. A number of authors (Dignum, Baldoni, and Baroglio, 2018; Kieslich *et al*., 2022; Craigon, Sacks, and Brewer, 2023) as well as governmental structures (European Commission, 2020) bring into debate the need to consider ethical aspects from the development phases of AI technologies.

Our research approaches the need for AI technology to be ethical by design from a different perspective. That's why the first component of our ethical framework is called Embedded Ethics By Design (AI EED). Our framework proposes through the AI EED component that every platform that can run or develop AI technologies have capabilities through which developers can train the models from the point of view of the ethical principles that must be respected.

This phase is considered to be a proactive way (of piloting the technology in question) that can be implemented by each organization before putting a particular AI system into production, taking into account the cultural context, the industry in which that AI technology is being implemented, the potential impact as well as who are the stakeholders involved in the subsequent management of the implemented system.

4.2 AI Desired Stated Configuration (AI DSC)

AI DSC is the component by which an AI technology is managed after its implementation, ensuring reliability and continuous updating of the underlying training models in different contexts with reference to ethical issues. Through the AI EED component of the proposed framework, it is possible to define and operationalize the ethical principles and values that a technology must incorporate at the algorithmic level (from the training model development phase). Post-implementation in production, an AI technology requires management that allows training models to be updated with new "scenarios" or "ethical challenges" that the technology can learn from new data obtained. As we showed in the Introduction section of this paper, validating the existence of ethics in AI technology decisions is not a one-time task or exercise. There is a need to ensure an ongoing process to validate, re-validate and refresh an AI technology post implementation. Our framework proposes three mechanisms (Monitoring, Evaluation, Mitigation) through which AI technologies can be ethically re-validated and updated once they have been implemented.

In the following sections we will focus the functional elements of the AI Ethical Validation Model component and the Monitoring component because these two are key differentiators for the proposed framework.

4.3 Functional aspects of the framework for the validation of ethical elements in Artificial Intelligence

In this section of the research, the relational elements of the ethical validation framework applicable to Artificial Intelligence will be described at the functional level. A series of conceptual definitions are required for a better interpretation of each element presented in the two components AI EED and AI DSC of the framework (according to Figure 5).

The AI Ethical Principle (PrinEticIA) - refers to moral values intended to help the development, implementation and responsible use of AI technologies.

AI Dasboard Interface (IntAdm IA) - refers to the graphical interface used to access the AI technology to be developed and trained.

Data Set (DS) - refer to categories of data that can be used at the algorithmic level in AI development.

Model Debugging - the training of an AI model trough a set of techniques and tools to analyse/interpret the training model from an ethical perspective.

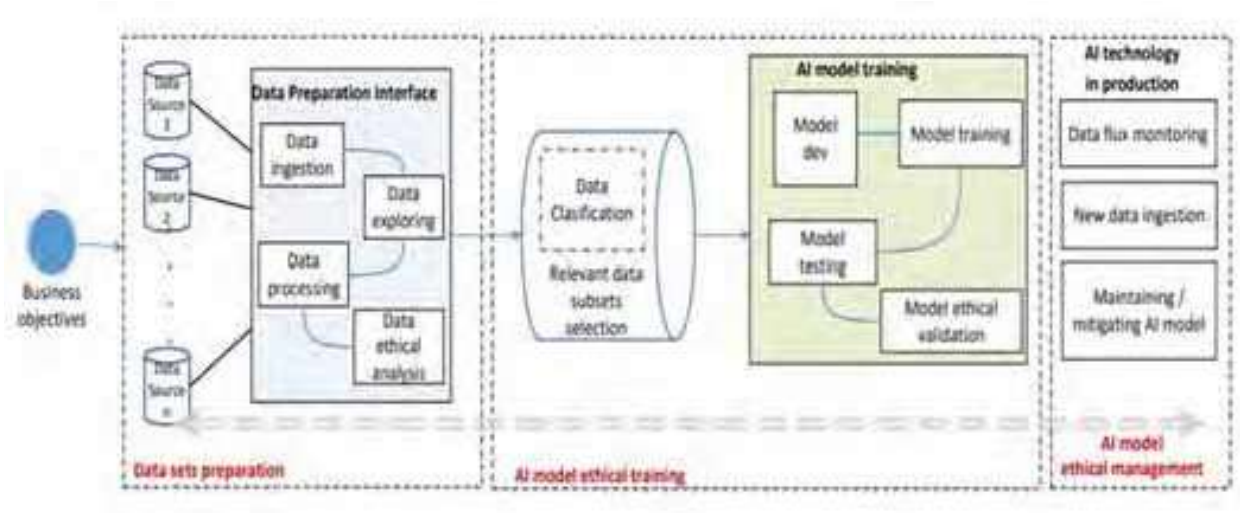
AI ethical validation model (AI EVM) - represents an analysis and decision stage, which includes two elements: model adjustment and a decision aiming to validate if an AI can be implemented under optimal conditions in production.

In the below sections we will describe the high-level architecture of the most important components of our frameworks, namely: the *DataSet*, the *AI Ethical Validation* model and the *AI DSC Monitoring* component. There are other functional elements in the framework that will be further described in future research papers.

4.3.1 The importance of DataSets for an AI ethical technology

The data set used to train the AI technology is a basic element, crucial to achieve the initially defined or anticipated ethical effects. Ideally, the data set should contain information as relevant as possible to the field in which the AI technology will be applied. Thus, it is extremely important that the data and data sources used by companies that want to implement AI technologies are analysed and verified to exclude possible ethical inconsistencies or biases that already exist at the data level.

Within the Figure 6 we are suggesting a method that is proposed by our framework and can be used to ethically identify and prepare data that is ingested in AI solutions.



Source: self-representation

Figure 6. Ingesting data in AI models while applying the ethical validation model

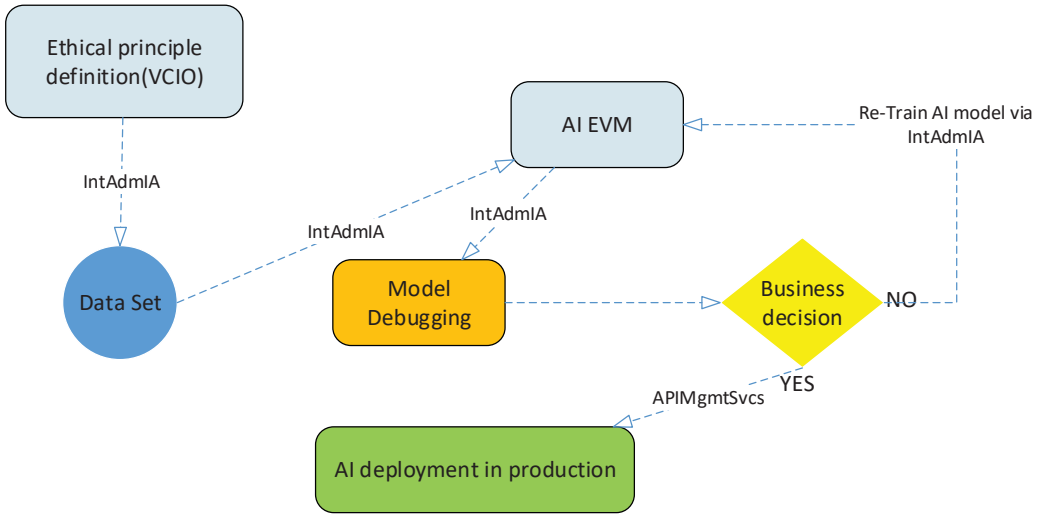
The process described in Figure 6 should be approached as a series of cyclical and interconnected activities and not as single-execution tasks to move to the next stage. For example, in the process of ethical training of the model, the lack of a certain category of data can be identified, which can lead to the identification of other sources of data to be prepared, classified, etc.

4.3.2 The AI ethical validation model of the proposed framework

Once the training model associated with the AI technology to be developed/implemented is created, an ethical evaluation of it is proposed in relation to the ethical principles defined by VCIO. This step is part of the AI EED component and is essential in ethically testing and validating AI technology before it is deployed in production. This stage in the framework proposes the creation of a process of ethical adjustment of the model (model debugging) whose results are correlated with a decision aimed at the initially defining ethical principles from the perspective of business objectives.

The ethical adjustment of the model (model debugging) consists in the use of those techniques that validate from a technical and functional point of view the decision-making mode of the AI technology. The framework proposes a list of open-source techniques already used by the industry (see Microsoft Responsible AI Dashboard) for the ethical validation of the training model. The list is not exhaustive and can always be adjusted according to the specific needs of the type of AI and its applicability.

In Figure 7 there's a schematic approach to the main stages within AI EVM that aim to train the AI model from an ethical point of view in order to implement AI technology in production at the level of business processes.



Source: Self-representation

Figure 7. Schematic approach to ethical validation in AI training model development

In the schema above, we represent a process of ethical validation at the level of development and testing of an AI training model, as follows:

Step 1 - Identify and define the ethical principle according to the VCIO model;

Step 2 - The relevant data sets are established;

Step 3 - The training model is built based on the identified data;

Step 4 - The training model will be ethically adjusted (if necessary);

Step 5 - Following the adjustment from step 4, the obtained input will be used, which will be correlated with the business-specific requirements. Then, a decision will be made to implement the AI in production.

Step 6 - If the decision is YES, the ethical validity of the training model is compliant with the business requirements, the AI technology will be implemented in production.

Step 7 - If the decision is NO, the training model is not considered ethical from the point of view of the business requirements, it goes back to step 3 to re-train it.

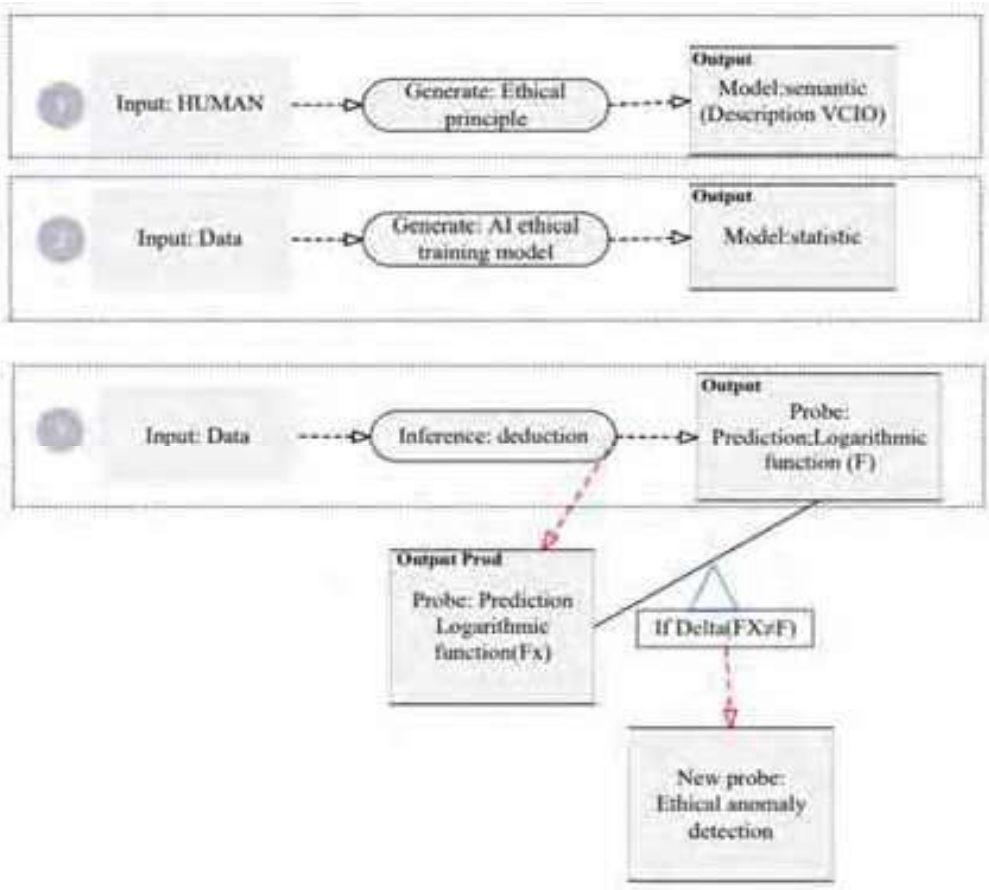
Between Step 4 and Step 5 there can be a two-way relationship in which AI practitioners technically test the model through the specific methods, communicate the results to the decision makers, so that they can propose new test scenarios based on specific business requirements.

4.3.3 The monitoring components of the AI ethical framework

Training models can certainly be tested a priori, using specific statistical and probabilistic methods, before they are released into production. However, the framework emphasizes the importance of monitoring and evaluation during the

use of AI. This mode has two major advantages: analytical reporting and traceability of possible ethical anomalies detection.

The framework, through the AI DSC component proposes an approach based on logical deductive reasoning for continuous ethical monitoring. In Figure 8 we propose the rational approach which is based on *input vs output* descriptive processes.



Source: Self-representation

Figure 8. The mechanism for detecting possible ethical anomalies post-implementation AI

Thus, a mechanism based on logical reasoning is proposed that starts from the following premises:

1. Need for a *Human* input to generate ethical principle(s) applicable to AI. *Output* will be a *semantic model* (definition of an ethical principle as per VCIO).
2. Translate from the semantic model to the logical, mathematical or statistical model, depending on the type of AI desired. Need for *Data input* to generate the basic AI training model. This process aims to obtain a *statistical representation output* (based on the initial relevant data sets).

3. AI ethical training and validation. *Input* is *Data* type, but by applying ethical training methods, the aim is to generate inferences based on *deductions*, after which a prediction-type output can result. These predictions are defined in the monitoring part of the framework as samples (**P**). Evidence is basically the sum of the predictions or decisions that are expected from an AI technology in relation to an initially defined scenario or ethical principle. Following the ethical training of the AI model will therefore result a sample that will contain predictions obtained through semantic, mathematical and statistical modeling that can be defined as logarithmic functions. As such, at this stage we will have predictions that can be defined logarithmically as **F(x).** This will be the sample that will work as a reference for the training model implemented in production.

Following the production deployment process, depending on new data or unanticipated scenarios during the training phase, the AI training model may make decisions and provide predictions that do not align with the reference sample **F(x).** In this case, a new sample of predictions will be deduced and mathematically defined **F(x)`**. Thus, the framework practically proposes the continuous monitoring and evaluation of the ratio between **F(x)** and **F(x)`** and the following assumptions:

I1 - If F(x) = F(x)`→ N (normal and expected AI ethical behavior)

- the behavior of the AI model and technology in production is ethically normal and no mitigation/adjustment is required. The originally trained ethical validity is practically confirmed.

I2 – If F(x) ≠ F(x)`→ Pn

- meaning there is a discrepancy between the expected predictions (initial sample) and the sample obtained in production (the new predictions), thus resulting in a *new sample* (**Pn**), which practically represents an potential ethical anomaly.

It should be noted that **Pn** will contain new data and information that formed the basis of the unexpected prediction. In this case *I2* should be tested by re-analyzing the data (including exploring any new data presented to the AI that led to the ethical anomaly). In order to validate *I2*, the ethical training methods established in AI EVM will be repeated to determine the cause of the anomaly. As such, the mitigation or adjustment process will begin, the framework proposing a mechanism for assimilating the new data and information that caused the ethical anomaly. It should be stated that there is also the possibility that a deviation is identified, but this is not necessarily an ethical anomaly. That is precisely why the framework's recommendation is to create a new customized data set (*DataLake type*) to retain the new data and only the ethically relevant ones to be re-assimilated into the initial training models. In this way, a separation will be

achieved between ethically relevant data for AI models and advisory data but without ethical relevance or impact.

5. CONCLUSIONS

This research paper aimed to address the question of how to establish a functional framework to ensure that AI technologies deployed in business environments make ethical decisions and produce ethical outcomes. The paper proposed an operational and comprehensive framework that can be used to validate the compliance of AI with ethical principles throughout its life cycle, from development to implementation and post-implementation in production. While the scientific literature contains numerous frameworks for the ethical validation of AI, most of them have limited practical operationalization of ethical principles in AI technologies.

The proposed framework offers an operational normalization that can translate ethical principles into a logarithmic level for practical implementation in AI technologies. The framework also emphasizes the traceability of the ethical validation process at the level of AI technologies, with two interoperable components AI EED and AI DSC that enable control over ethical principles from the perspective of business objectives and in the implementation and post-implementation phase of AI.

The proposed framework offers visibility and control over ethical validation, which is useful in situations such as computer audits. Additionally, the development of a national/international AI ethical control body that uses the proposed framework as a tool for verifying ethical norms applied in AI could be a direction for further development.

Lastly, the AI DSC monitoring and detection component of the framework also has an analytical role in providing specific information for businesses on the use of the implemented AI solution, the impact of ethical violations, and the degree of prediction of AI. This information can lead to a strategic advantage that improves the relationship between business, technology, and added value, with positive social impact. Overall, the proposed framework provides a valuable tool for ethical validation of AI technologies in business environments.

References

1) AI Ethics Group Impact (2020). [online] Available at: www.ai-ethics-impact.org [Accessed 15.03.2023].
2) Algorithm Watch (2023). [online] Available at: https://algorithmwatch.org/en/ [Accessed 15.03.2023].
3) Alshammari, M. and Simpson, A. (2017). Towards a Principled Approach for Engineering Privacy by Design. *Annual Privacy Forum.* DOI:10.1007/978-3-319-67280-9_9.

4) Anderson, S. and Anderson, M. (2021). AI and ethics. *AI Ethics,* 1, pp. 27–31. https://doi.org/10.1007/s43681-020-00003-6.
5) Ayling, J. and Chapman, A. (2022). Putting AI ethics to work: are the tools fit for purpose? *AI Ethic,* 2, p. 405–429. https://doi.org/10.1007/s43681-021-00084-x.
6) Burt, A. (2020). *Ethical Frameworks for AI Aren't Enough.* [online] Available at: https://hbr.org/2020/11/ethical-frameworks-for-ai-arent-enough. [Accessed 15.03.2023].
7) Calvo, R. and Peters, D. (2019). *Design for Wellbeing - Tools for Research, Practice and Ethics.* New York, NY, In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19). Association for Computing Machinery.
8) Ciobanu, C. A. and Meșniță, G. (2022). *AI Ethics for Industry 5.0 – From Principles to Practice.* Valencia, Spain, CEUR Workshop Proceedings (CEUR-WS.org).
9) Craigon, P., Sacks, J. and Brewer, S. (2023). Ethics by design: Responsible research and innovation for AI in the food sector. *Journal of Responsible Technology,* 13(100051). https://doi.org/10.1016/j.jrt.2022.100051.
10) Dignum, V., Baldoni, M. and Baroglio, C. (2018). *Ethics by Design: Necessity or Curse? In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18).* New York, Association for Computing Machinery, pp. 60–66.
11) Europenan Commission (2021). *European Commission.* [online] Available at: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence. [Accessed 25.01.2023].
12) Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence,* 1,pp. 261-262. https://doi.org/10.1038/s42256-019-0055-y.
13) Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P. and Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines,* 28(4), pp. 689-707. DOI: 10.1007/s11023-018-9482-5.
14) Haas, L., Gießler, S. and Thiel, V. (2020). *In the realm of paper tigers – exploring the failings of AI ethics guidelines.* [online] Available at: https://algorithmwatch.org/en/ai-ethics-guidelines-inventory-upgrade-2020/. [Accessed 20.11. 2022].
15) Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines,* 30(1), pp. 99-120.
16) Hallensleben, S. and Hustdet, C. (2020). *From Principles to Practice:An interdisciplinary framework to operationalise AI ethics.* [online] Available at: https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf 2019 [Accessed 15.02.2022].
17) Jobin, A., Ienca, M. and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence,* 1(9), pp. 389-399.
18) Kieslich, K., Keller, B. and Starke, C. (2022). Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence. *Big Data and Society,* 9(1). https://doi.org/10.1177/205395172210929.

19) Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nat Mach Intell,* 1, pp. 501–507. DOI: 10.1038/s42256-019-0114-4.
20) Morley, J., Elhalal, A. and Garcia, F. (2021). Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. *Minds and Machines,* 31, pp. 239–256. https://doi.org/10.1007/s11023-021-09563-w.
21) Morley, J., Floridi, L., Kinsey, L. and Elhalal, A. (2020). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics,* 26(3), pp. 2141–2168. DOI: 10.1007/s11948-019-00165-5.
22) Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M. and Floridi, L. (2023). Operationalising AI ethics: barriers, enablers and next steps.. *AI & Society ,* 38, pp. 411–423. https://doi.org/10.1007/s00146-021-01308-8.
23) Stephanidis, C. (2019). Seven HCI Grand Challenges. *International Journal of Human-Computer Interaction,* 35(14), pp. 1229-1269. DOI: 10.1080/10447318.2019.1619259.
24) Turilli, M. (2007). Ethical protocols design. *Ethics and Information Technology,* 9(1), pp. 49–62.
25) Turing, A. (1950). Computing Machinery and Intelligence. *Mind,* LIX(236), pp. 433–460.
26) Vidgen, R., Hindle, G. and Randolph, I. (2020). Exploring the ethical implications of business analytics with a business ethics canvas. *European Journal of Operational Research,* 281(3), pp. 491-501.
27) Ville, V., Kai-Kristian, K. and Pekka, A. (2019). *AI ethics in industry: A research framework.* [online] Available at: https://arxiv.org/abs/1910.12695. [Accessed 18.02.2023].
28) Wiener, N. (1954). *The human use of human beings.* New York: Doubleday.